ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CNTT&TT

Nguyễn Thị Sinh

KỸ THUẬT PHÂN CỤM DỮ LIỆU ỨNG DỤNG TRONG GIS

Chuyên ngành: Khoa học máy tính

Mã số: 60 48 01 01

LUẬN VĂN THẠC SĨ CHUYÊN NGÀNH KHOA HỌC MÁY TÍNH

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

Thái Nguyên - 2014

MỞ ĐẦU

Khai phá dữ liệu không gian hay còn gọi là khai phá tri thức từ dữ liệu không gian là một lĩnh vực có nhu cầu rất cao. Bởi lẽ dữ liệu đầu vào ở đây bao gồm một khối lượng dữ liệu không gian khổng lồ đã được thu thập từ nhiều ứng dụng khác nhau, 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.

Trong những năm gần đây, việc nghiên cứu về khai phá dữ liệu đã có xu hướng chuyển từ cơ sở dữ liệu quan hệ và cơ sở dữ liệu giao dịch sang cơ sở dữ liệu không gian. Sự thay đổi này không những giúp hiểu được dữ liệu không gian mà còn giúp khám phá được mối quan hệ giữa dữ liệu không gian và phi không gian, các mô hình dựa trên tri thức không gian, phương pháp tối ưu câu truy vấn, tổ chức dữ liệu trong cơ sở dữ liệu không gian, ... 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, 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 Nhà thi đấu Hải Dương ?
- Những căn nhà nào nằm trong vùng quy hoạch mở rộng phố?

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

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

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

xác định được vị trí đặt các trạm bán lẻ xăng dầu cần phải tuân theo các quy định của Bộ Công thương, nhất là các quy định về an toàn, phòng chống cháy nổ. Ngoài ra, cây xăng cũng phải đặt ở vị trí thuận lợi cho việc kinh doanh đạt doanh số cao.

Hoặc một bài toán khác cũng có ý nghĩa thực tiễn rất lớn đó là xác định vị trí tối ưu để mở một nhà hàng. Hiện nay trên địa bàn thành phố Hà Nội cũng đã có rất nhiều nhà hàng, quán ăn đã được mở ra. Nhưng không phải tất cả các nhà hàng, quán ăn đó đều có thể cho doanh thu tốt. Có khi có nhà hàng mới mở ra được một thời gian ngắn đã phải đóng cửa vì không có khách dẫn đến chủ đầu tư phải chịu thua lỗ nặng. Một trong những nguyên nhân chính dẫn đến thất bại đó là địa điểm kinh doanh chưa hợp lý. Một vị trí tối ưu cho việc mở nhà hàng, quán ăn thì vị trí đó phải thỏa mãn một số yếu tố sau: nằm trong khu vực đông dân cư, gần nhiều cơ quan công sở hay trường học, có khu vực để xe, có quang cảnh xung quanh thoáng mát...

Xuất phát từ nhu cầu thực tế đó, luận văn giới thiệu một số phương pháp phân cụm dữ liệu trong khai phá cơ sở dữ liệu không gian 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, trong đó khai thác thông tin địa lý của các đối tượng địa lý để hỗ trợ giải quyết bài toán ví dụ như tìm vị trí tối ưu đặt nhà hàng hoặc các trạm xăng dầu trong thành phố Hà Nội.

Luận văn được chia thành các chương mục sau:

- Mở đầu
- Chương 1: Tổng quan về Hệ thông tin Địa lý (GIS)
- Chương 2: Kỹ thuật phân cụm dữ liệu không gian
- Chương 3: Xây dựng chương trình thử nghiệm Kết luận, đánh giá
- Kết luận

CHƯƠNG 1: TỔNG QUAN VỀ HỆ THÔNG TIN ĐỊA LÝ (GIS)

1.1 Mô hình dữ liệu địa lý:

Khái niệm Địa lý (*Geography*) đề cập lĩnh vực nghiên cứu mô tả Trái đất (*Geo-Earth*). Ngày nay, khái niệm này và khái niệm Không gian (*Space*) được sử dụng thay thế nhau trong một số trường hợp. Tuy nhiên, về mặt bản chất thì Địa lý là tập các mô tả về không gian (hai chiều), khí quyển (ba chiều), … của Trái đất. Còn không gian cho phép mô tả bất kỳ cấu trúc đa chiều nào, không quan tâm đến vị trí địa lý của nó. Như vậy có thể coi Địa lý như là một phần cấu trúc nhỏ trong tập cấu trúc Không gian.

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 cách diễn giải khác nhau cho từ viết tắt GIS, tuy nhiên các cách diễn giải đó 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 Biểu diễn dữ liệu địa lý

Các thành phần của dữ liệu địa lý

Trong GIS, dữ liệu được chia làm hai loại: thành phần không gian và thành phần phi không gian (thuộc tính). Hai loại thành phần dữ liệu này được kết hợp thông qua một chỉ số chung để mô tả một đối tượng thực. Sự kết hợp này thể hiện đặc trưng không gian của đối tượng, nó cho phép:

* Mô tả *"vị trí, hình dạng"*: vị trí tham chiếu, đơn vị đo, dạng hình học của thực thể địa lý.

* Mô tả *"quan hệ và tương tác"* giữa các thực thể địa lý: những thửa đất nào liền kề với khu công nghiệp ?

* Mô tả "*thông tin*" của các đối tượng địa lý: ai là chủ sở hữu của thửa đất này?

Thành phần không gian

Thành phần dữ liệu không gian hay còn gọi là 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:

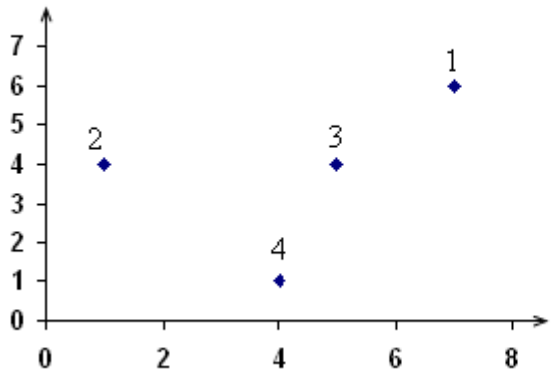
- **Điểm (*Point*)**

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 1.1 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

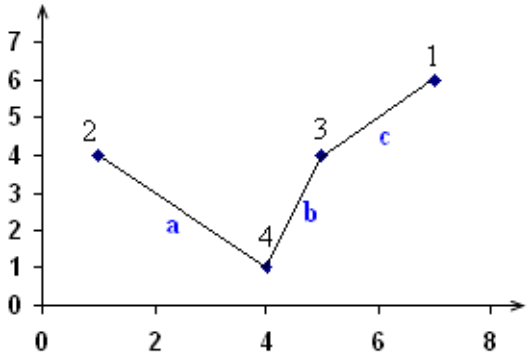
DiemID	Độ sâu	Lượng nước
1	12.5	63.5
2	26	80
3	3.5	24.34
4	17	69.23

Hình 1.1: 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

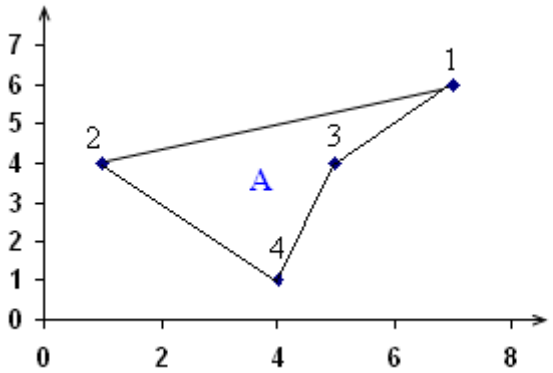
DuongID	Lưu lượng	Dung lượng
a	960	2220
b	1250	2000
c	1100	1950

Hình 1.2: 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 1.3 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	2.5

Hình 1.3: 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.

1.1.3 Mô hình biểu diễn dữ liệu địa không gian

Như đã đề cập ở trên, dữ liệu địa lý bao gồm thành phần dữ liệu không gian và thành phần dữ liệu thuộc tính. Ở phần này, chúng ta sẽ xem xét cách thức biểu diễn thành phần dữ liệu không gian trong hệ thông tin địa lý.

Hệ thông tin địa lý biểu diễn các thực thể địa lý trong tự nhiên bằng dữ liệu của nó, hệ thống GIS chứa càng nhiều dữ liệu thì khả năng mang lại thông tin càng lớn. 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 quan 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