ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

BÙI THỊ HUẾ

KHAI PHÁ DỮ LIỆU VỚI HỆ THÔNG TIN ĐỊA LÝ VÀ ỨNG DỤNG

LUẬN VĂN THẠC SĨ: KHOA HỌC MÁY TÍNH

Thái Nguyên – 2013

MỞ ĐẦU

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

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

Là giáo viên của tỉnh Nam Định, em đã tích lũy được khá nhiều dữ liệu thực tế về xếp hạng các trường THPT trong tỉnh qua kết quả thi tốt nghiệp, kết quả thi đại học và kết quả thi học sinh giỏi,... Trong đó, rất nhiều trường có thành tích cao nhưng cũng tồn tại không ít các trường có kết quả học tập của học sinh còn rất thấp. Nguyên nhân nào dẫn đến kết quả đó? Liệu điều kiện nơi cư trú có ảnh hưởng đến kết quả học tập của học sinh? Lượng thông tin về kết quả học tập và điều kiện cư trú ở mỗi địa phương rất nhiều và chủ yếu quản lý ở dạng bảng hay văn bản. Rất khó để thể hiện mối liên hệ giữa điều kiện cư trú với kết quả học tập của học sinh theo hướng khai phá dữ liệu thông thường.

Ứng dụng khai phá dữ liệu với hệ thông tin địa lý cho phép nghiên cứu các vấn đề, trả lời các câu hỏi có liên quan trực tiếp, ví dụ như “ô nhiễm môi trường sống ảnh hưởng như thế nào đến sức khỏe cộng đồng” và cả những câu

hỏi ít trực tiếp hơn, ví dụ như "nơi cư trú của học sinh (ở đô thị lớn, ở nông thôn, ở vùng núi,…) có ảnh hưởng như thế nào đến kết quả học tập của những môn học nhất định (về khoa học tự nhiên, khoa học xã hội, ngoại ngữ,...) đến số học sinh bỏ học, đến trung bình kết quả học tập, đến số học sinh đỗ đại học, số học sinh giỏi?...".

Luận văn đặt vấn đề ứng dụng khai phá dữ liệu không gian với hệ thông tin địa lý để tìm hiểumối liên hệ giữa nơi cư trú và kết quả học tập với mục tiêu bước đầu thử nghiệm áp dụng một số kỹ thuật khai phá dữ liệu thường dùng với GIS vào bài toán thực tế.

Luận văn cấu trúc gồm 3 chương:

Chương I: Trong chương 1 sẽ tìm hiểu khái quát về khai phá dữ liệu và khai phá dữ liệu không gian.

Chương II: Trong chương 2 sẽ tìm hiểu một số thuật toán phân cụm và kỹ thuật phân cụm bằng thuật toán CLARANS.

Chương III: Trong chương 3 tiến hành cài đặt ứng dụng thuật toán CLARANS để phân cụm dữ liệu không gian, tìm hiểu mối liên hệ giữa điều kiện cư trú với kết quả học tập của học sinh.

Luận văn này được hoàn thành dưới sự hướng dẫn tận tình của PGS.TS **Nguyễn Đình Hóa**, em xin bày tỏ lòng biết ơn chân thành của mình đối với thầy. Em xin chân thành cảm ơn các thầy, cô giáo Viện Công nghệ thông tin, Trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên đã tham gia giảng dạy, giúp đỡ em trong suốt qúa trình học tập nâng cao trình độ kiến thức. Tuy nhiên vì điều kiện thời gian và khả năng có hạn nên luận văn không thể tránh khỏi những thiếu sót. Em kính mong các thầy cô giáo và các bạn đóng góp ý kiến để đề tài được hoàn thiện hơn.

CHƯƠNG I: KHAI PHÁ DỮ LIỆU VỚI HỆ THÔNG TIN ĐỊA LÝ

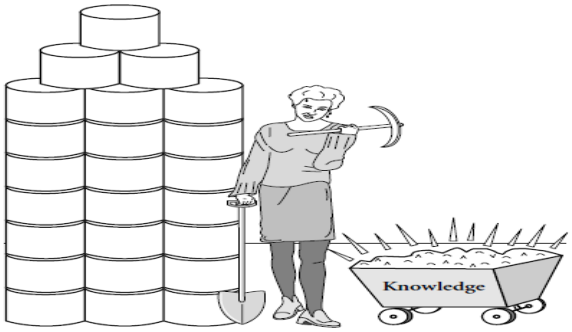
1.1 Khai phá dữ liệu

1.1.1 Khai phá dữ liệu là gì ?

1.1.1.1 Khái niệm về khám phá tri thức và khai phá dữ liệu

Khám phá tri thức trong các cơ sở dữ liệu (*Knowledge Discovery in Database-KDD***)** là một qui trình nhận biết các mẫu hoặc các mô hình trong dữ liệu với các tính năng: hợp thức, mới, khả ích, và có thể hiểu được.

Khai phá dữ liệu (Data Mining-DM) là một khái niệm ra đời vào những năm cuối của thập kỷ 1980. Cụm từ "***khai phá dữ liệu"*** bao hàm một loạt các kỹ thuật nhằm phát hiện ra các thông tin có giá trị tiềm ẩn trong các tập dữ liệu lớn. Thuật ngữ này thực sự là một cái tên nhầm lẫn. Hãy nhớ rằng việc khai thác vàng từ đá hoặc cát được gọi là khai thác vàng chứ không phải là khai thác đá, cát. Như vậy, khai phá dữ liệu (KPDL) nên được đặt tên thích hợp hơn là "***khai thác kiến thức từ dữ liệu***" [5].Tuy nhiên, "khai phá dữ liệu" vẫn được dùng cách phổ biến. Hình 1.1 minh họa đơn giản và trực quan cho khái niệm này.



Hình 1.1: Khai phá dữ liệu trong tập dữ liệu [5]

Khái niệm KDD và KPDL được các nhà khoa học xem là tương đương nhau. Tuy nhưng, nếu phân chia một cách rành mạch và chi tiết thì KPDL là một bước chính trong quá trình KDD.

Khám phá tri thức trong CSDL là lĩnh vực liên quan đến nhiều ngành như: Tổ chức dữ liệu, xác suất, thống kê, lý thuyết thông tin, học máy, CSDL, thuật toán, trí tuệ nhân tạo, tính toán song song và hiệu năng cao. Các kỹ thuật chính áp dụng trong khám phá tri thức phần lớn được thừa kế từ các ngành này.

1.1.1.2 Một số định nghĩa về khai phá dữ liệu

Sau đây là một số định nghĩa khác nhau về KPDL [5]:

Định nghĩa của Giáo sư **Tom Mitchell:** "*Khai phá dữ liệu là việc sử dụng dữ liệu lịch sử để khám phá những qui tắc và cải thiện những quyết định trong tương lai*."

Định nghĩa của **Ferruzza**: "*Khai phá dữ liệu là tập hợp các phương pháp được dùng trong tiến trình khám phá tri thức để chỉ ra sự khác biệt các mối quan hệ và các mẫu chưa biết bên trong dữ liệu*"

Định nghĩa của **Parsaye**: "*Khai phá dữ liệu là quá trình trợ giúp quyết định, trong đó chúng ta tìm kiếm các mẫu thông tin chưa biết và bất ngờ trong CSDL lớn*"

Với một cách tiếp cận ứng dụng hơn, **tiến sĩ Fayyad** đã phát biểu:" *Khai phá dữ liệu thường được xem là việc khám phá tri thức trong các cơ sở dữ liệu, là một quá trình trích xuất những thông tin ẩn, trước đây chưa biết và có khả năng hữu ích, dưới dạng các quy luật, ràng buộc, qui tắc trong cơ sở dữ liệu*."

Còn các nhà **Thống kê** thì xem *"Khai phá dữ liệu như là một quá trình phân tích được thiết kế thăm dò một lượng cực lớn các dữ liệu nhằm phát hiện ra các mẫu thích hợp và/ hoặc các mối quan hệ mang tính hệ thống giữa các biến và sau đó sẽ hợp thức hoá các kết quả tìm được bằng cách áp dụng các mẫu đã phát hiện được cho tập con mới của dữ liệu*".

Tuy nhiên trên thực tế, KPDL được xem là một bước thiết yếu trong quá trình khám phá tri thức trong CSDL bao gồm các thụật toán KPDL chuyên dùng, dưới một số quy định về hiệu quả tính toán chấp nhận được, để tìm ra các mẫu hoặc các mô hình trong dữ liệu.

1.1.2 Quá trình phát hiện tri thức trong CSDL

Quá trình phát hiện tri thức trong CSDL được mô tả trong hình 1.2 và bao gồm một chuỗi lặp đi lặp lại các bước sau [4] :

Làm sạch dữ liệu (Data Cleaning): Loại bỏ nhiễu (noisy) và các dữ liệu không nhất quán.

Tích hợp dữ liệu (Data Integration): Kết hợp dữ liệu từ các nguồn dữ liệu khác nhau.

Lựa chọn dữ liệu (Data Selection): Dữ liệu phù hợp cho thao tác phân tích được lấy về từ cơ sở dữ liệu.

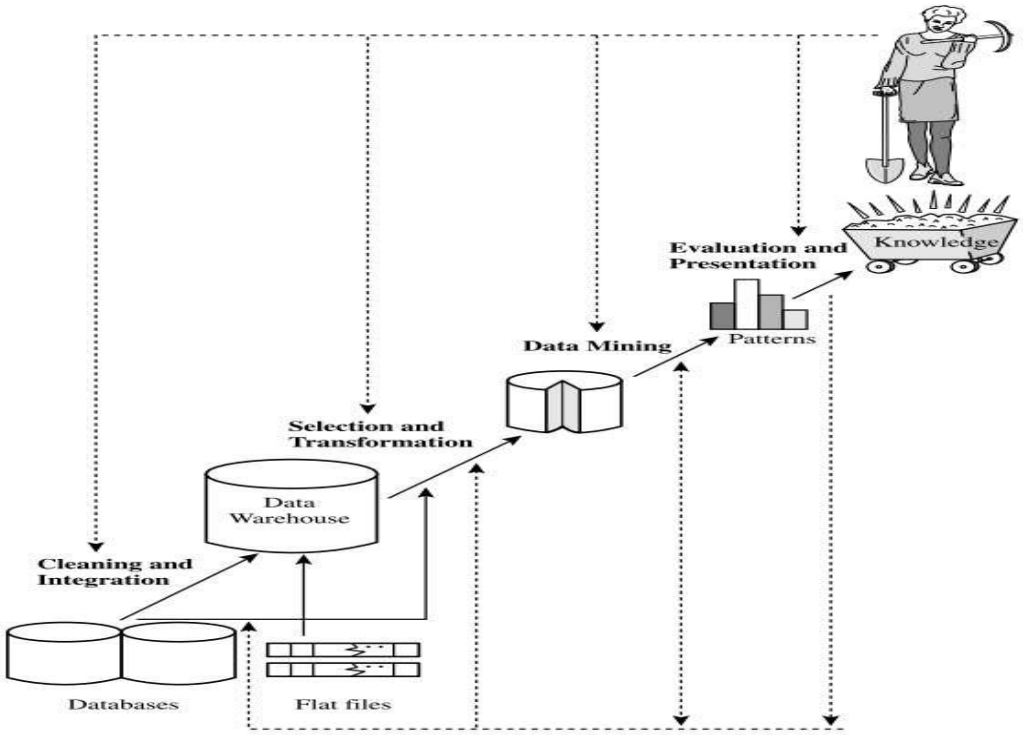
Chuyển dạng dữ liệu (Data Transformation): Dữ liệu được chuyển dạng hoặc hợp nhất thành những dạng phù hợp cho quá trình khai phá bằng cách thực hiện các thao tác như tóm tắt (summary) hoặc gộp nhóm dữ liệu (aggregation).

Trích chọn mẫu (Data Patterns Extracting): Áp dụng các phương pháp "thông minh" để trích chọn ra các mẫu thực sự đáng quan tâm từ dữ liệu. Đôi khi chính bản thân bước này cũng được gọi là khai phá dữ liệu (Data Mining) (hiểu theo nghĩa hẹp).

Đánh giá mẫu (Pattern Evaluation): Dựa trên các độ đo đặc trưng, xác định ra các mẫu đáng quan tâm biểu diễn tri thức.

Biểu diễn tri thức (Knowledge Presentation): Sử dụng các kỹ thuật biểu diễn tri thức và trực quan hóa (visualization) để biểu diễn và biến đổi các tri

thức khai phá được ở bước trên thành một dạng gần gũi với người sử dụng như đồ thị, cây, bảng biểu, luật,... đến với người dùng.



Hình 1.2: Quy trình khám phá tri thức từ cơ sở dữ liệu [4]

Trong đó, 4 giai đoạn đầu được gọi là các giai đoạn tiền xử lý dữ liệu (data preprocessing) nhằm chuẩn bị dữ liệu cho quá trình khai phá (trích chọn mẫu).

Các giai đoạn của quá trình khám phá tri thức có mối quan hệ chặt chẽ với nhau trong bối cảnh chung của hệ thống. Các kỹ thuật được sử dụng trong giai đoạn trước có thể ảnh hưởng đến hiệu quả của các giải thuật được sử dụng trong các giai đoạn tiếp theo. Quá trình KDD không nhất thiết phải tuần tự, nó cho phép các nhà phân tích có thể xem xét lại các bước dựa trên những kiến thức tìm kiếm và bản chất của các thông tin được phát hiện trong quá trình. Các bước tiền xử lý dữ liệu như chế biến làm sạch, lựa chọn và rút gọn có thể được áp dụng theo các trình tự khác nhau và có thể lặp đi lặp lại một số lần.

1.1.3 Các kỹ thuật khai phá dữ liệu

Trong thực tế có nhiều kỹ thuật khai phá dữ liệu khác nhau nhằm thực hiện hai chức năng mô tả và dự đoán.

Kỹ thuật khai phá dữ liệu mô tả: có nhiệm vụ mô tả các tính chất hoặc các đặc tính chung của dữ liệu trong CSDL hiện có. Một số kỹ thuật khai phá trong nhóm này là: phân cụm dữ liệu (Clustering), tổng hợp (Summarisation), trực quan hoá (Visualization), phân tích sự tiến hóa (Evolution and deviation analyst),….

Kỹ thuật khai phá dữ liệu dự đoán: có nhiệm vụ đưa ra các dự đoán dựa vào các suy diễn trên cơ sở dữ liệu hiện thời. Một số kỹ thuật khai phá trong nhóm này là: phân lớp (Classification), hồi quy (Regression), cây quyết định (Decision tree), thống kê (statictics), mạng nơron (neural network), luật kết hợp,….

Một số kỹ thuật phổ biến [1],[3],[5] thường được sử dụng để khai phá dữ liệu hiện nay là :

1.1.3.1 Phân lớp dữ liệu

Mục tiêu của phân lớp dữ liệu đó là dự đoán nhãn lớp cho các mẫu dữ liệu. Quá trình gồm hai bước: xây dựng mô hình, sử dụng mô hình để phân lớp dữ liệu (mỗi mẫu 1 lớp). Mô hình được sử dụng để dự đoán nhãn lớp khi mà độ chính xác của mô hình chấp nhận được.

1.1.3.2 Phân cụm dữ liệu

Mục tiêu của phân cụm dữ liệu là nhóm các đối tượng tương tự nhau trong tập dữ liệu vào các cụm, sao cho các đối tượng thuộc cùng một cụm là tương đồng.

Trong luận này tác giả đã sử dụng kỹ thuật phân cụm và thuật toán CLARANS tìm hiểu mối liên hệ giữa điều kiện cư trú với kết quả học tập của

học sinh. Vì vậy kỹ thuật này và các thuật toán có liên quan sẽ được trình bày trong chương II.

1.1.3.3 Khai phá luật kết hợp

Mục tiêu của phương pháp này là phát hiện và đưa ra các mối liên hệ giữa các giá trị dữ liệu trong cơ sở dữ liệu. Đầu ra của giải thuật luật kết hợp là tập luật kết hợp tìm được. Phương pháp khai phá luật kết hợp gồm có hai bước:

Bước 1: Tìm ra tất cả các tập mục phổ biến. Một tập mục phổ biến được xác định thông qua tính độ hỗ trợ và thoả mãn độ hỗ trợ cực tiểu.

Bước 2: Sinh ra các luật kết hợp mạnh từ tập mục phổ biến, các luật phải thoả mãn độ hỗ trợ và độ tin cậy cực tiểu.

1.1.3.4 Hồi quy

Phương pháp hồi quy tương tự như là phân lớp dữ liệu. Nhưng khác ở chỗ nó dùng để dự đoán các giá trị liên tục còn phân lớp dữ liệu dùng để dự đoán các giá trị rời rạc.

1.1.3.5 Mạng nơ-ron (neural network)

Đây là một trong những kỹ thuật KPDL được ứng dụng phổ biến hiện nay. Kỹ thuật này phát triển dựa trên một nền tảng toán học vững vàng, khả năng huấn luyện trong kỹ thuật này dựa trên mô hình thần kinh trung ương của con người.

Kết quả mà mạng nơ-ron học được có khả năng tạo ra các mô hình dự báo, dự đoán với độ chính xác và độ tin cậy cao. Nó có khả năng phát hiện ra được các xu hướng phức tạp mà kỹ thuật thông thường khác khó có thể phát hiện ra được. Tuy nhiên phương pháp neural network rất phức tạp và quá trình tiến hành nó gặp rất nhiều khó khăn: đòi hỏi mất nhiều thời gian, nhiều DL, nhiều lần kiểm tra thử nghiệm.

1.1.3.6 Cây quyết định

Kỹ thuật cây quyết định là một công cụ mạnh và hiệu quả trong việc phân lớp và dự báo. Các đối tượng DL được phân thành các lớp. Các giá trị của đối tượng DL chưa biết sẽ được dự đoán, dự báo. Tri thức được rút ra trong kỹ thuật này thường được mô tả dưới dạng tường minh, đơn giản, trực quan, dễ hiểu đối với người sử dụng. Trong những năm qua, nhiều mô hình phân lớp DL đã được các nhà khoa học trong nhiều lĩnh vực khác nhau đề xuất, nhưng kỹ thuật cây quyết định với những ưu điểm của mình được đánh giá là một công cụ mạnh, phổ biến và đặc biệt thích hợp cho DM nói chung và phân lớp dữ liệu nói riêng.

1.2 Khai phá dữ liệu GIS

1.2.1 Khái niệm

Dữ liệu địa lý đồ sộ đã và sẽ tiếp tục được thu thập bởi công nghệ thu thập dữ liệu hiện đại như hệ thống định vị toàn cầu (GPS), độ phân giải cảm biến từ xa, dịch vụ địa điểm nhận biết vị trí, các cuộc điều tra, và thông tin địa lý tình nguyện trên internet. Có một nhu cầu cấp thiết cho các phương pháp hiệu chính kịp thời và hiệu quả để trích xuất thông tin tiềm ẩn và bất ngờ từ bộ dữ liệu không gian rộng lớn và độ phức tạp cao. Để giải quyết những thách thức này, khai thác dữ liệu không gian và khám phá tri thức địa lý đã nổi lên như một lĩnh vực nghiên cứu hoạt động, tập trung vào sự phát triển của lý thuyết, phương pháp và thực hành cho việc khai thác các thông tin hữu ích và kiến thức từ cơ sở dữ liệu không gian lớn và phức tạp [6], [9].

Khai phá dữ liệu với GIS hay cũng gọi là khai phá dữ liệu không gian, mở rộng khai phá dữ liệu trong các CSDL quan hệ, xét thêm các thuộc tính của dữ liệu không gian được phản ánh trong hệ thông tin địa lý.

Phương pháp khai phá dữ liệu thông thường có thể không phù hợp với dữ liệu không gian bởi vì chúng không hỗ trợ các dữ liệu về vị trí địa lý cũng như mối quan hệ tiềm ẩn giữa các đối tượng địa lý. Do đó, cần phát triển các