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

---------- ---------- ----------

PHẠM THỊ LÝ

Tên đề tài:

KHAI PHÁ TẬP MỤC THƯỜNG XUYÊN ĐÓNG TRÊN DÒNG DỮ LIỆU

Chuyên ngành: KHOA HỌC MÁY TÍNH
Mã số : 60.48.01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Hướng dẫn khoa học: TS. NGUYỄN HUY ĐỨC

Thái Nguyên - 2014

MỞ ĐẦU

Khai phá dữ liệu (Data Mining), hiện nay đang được rất nhiều người chú ý. Nó thực sự đã đem lại những lợi ích đáng kể trong việc cung cấp những thông tin tiềm ẩn trong các cơ sở dữ liệu lớn, giúp người sử dụng thu được những tri thức hữu ích từ những cơ sở dữ liệu hoặc các kho dữ liệu khổng lồ khác. Những "tri thức" chiết xuất từ nguồn cơ sở dữ liệu đó phục vụ các yêu cầu trợ giúp quyết định ngày càng có ý nghĩa quan trọng và là nhu cầu to lớn trong mọi lĩnh vực hoạt động kinh doanh, quản lý. Tiến hành công việc như vậy chính là thực hiện quá trình phát triển tri thức trong cơ sở dữ liệu (Knowledge Discovery in Database) mà trong đó kỹ thuật khai phá dữ liệu (Data Mining) cho phép phát hiện những tri thức tiềm ẩn.

Một trong các nội dung cơ bản trong khai phá dữ liệu là khai phá luật kết hợp. Khai phá luật kết hợp gồm hai bước: Bước thứ nhất, tìm tất cả các tập mục thường xuyên, đòi hỏi sự tính toán lớn. Bước thứ hai, dựa vào các tập mục thường xuyên tìm các luật kết hợp, đòi hỏi tính toán ít hơn, song gặp phải một vấn đề là có thể sinh ra quá nhiều luật, vượt khỏi sự kiểm soát của người khai phá hoặc người dùng, trong đó có nhiều luật không cần thiết. Để giải quyết vấn đề đó, trong bước thứ nhất, không cần thiết khai phá tất cả các tập mục thường xuyên mà chỉ cần khai phá các tập mục thường xuyên đóng. Khai phá luật kết hợp dựa trên tập mục thường xuyên đóng cho hiệu quả cao hơn, nó đảm bảo không tìm ra các tập mục thường xuyên không cần thiết, không sinh ra các luật dư thừa.Với ý nghĩa đó và mục đích tìm hiểu về bài toán tìm tập mục thường xuyên trên dòng dữ liệu, em đã quyết định lựa chọn đề tài "Khai phá tập mục thường xuyên đóng trên dòng dữ liệu".

Nội dung luận văn gồm 3 chương:

Chương 1: Tổng quan về khai phá dữ liệu

Chương 2: Khai phá tập mục thường xuyên đóng trên dòng dữ liệu

Chương 3: Chương trình thực nghiệm ứng dụng

CHƯƠNG 1 TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

1.1. Khái niệm về khám phá tri thức và khai phá dữ liệu.

KPDL (Khai phá dữ liệu) là việc rút trích tri thức một cách tự động và hiệu quả từ một khối dữ liệu lớn. Tri thức đó thường ở dạng các mẫu có tính chất không tầm thường, không tường minh (ẩn), chưa được biết đến và có tiềm năng mang lại lợi ích. Có một số nhà nghiên cứu còn gọi khai phá dữ liệu là phát hiện tri thức trong cơ sở dữ liệu (Knowledge Discovery in Database - KDD). Ở đây chúng ta có thể coi KPDL là cốt lõi của quá trình phát hiện tri thức. Quá trình phát hiện tri thức gồm các bước sau :

Bước 1: Trích chọn dữ liệu (Data Selection). Là bước trích chọn những tập dữ liệu cần được khai phá từ các tập dữ liệu lớn (databases, data warehouses).

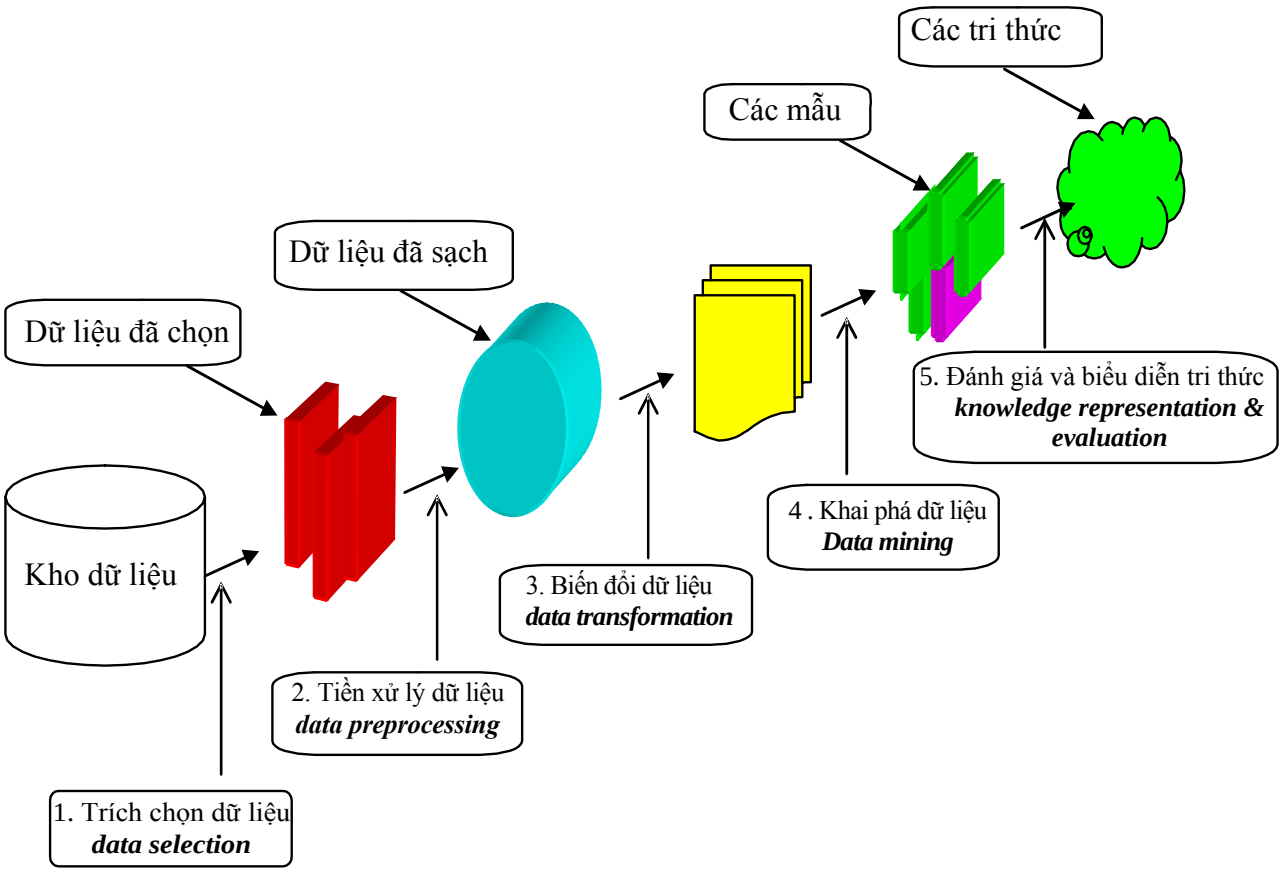
Bước 2: Tiền xử lý dữ liệu (Data preprocessing) là bước làm sạch dữ liệu (Xử lý dữ liệu không đầy đủ, dữ liệu nhiễu, dữ liệu không nhất quán…rút gọn dữ liệu (Sử dụng các phương pháp thu gọn dữ liệu, histograms, lấy mẫu, v..v..) rời rạc hóa dữ liệu (dựa vào histograms, entropy, phân khoảng.v..v..). Sau bước này dữ liệu sẽ nhất quán, đầy đủ, được rút gọn và được rời rạc hóa.

Bước 3: Biến đổi dữ liệu (Data transformation) Là bước chuẩn hóa và làm mịn dữ liệu để đưa dữ liệu về dạng thuận lợi nhất nhằm phục vụ cho các kỹ thuật khai thác ở bước sau.

Bước 4 : Khai phá dữ liệu (Data mining) Đây là bước quan trọng và tốn nhiều thời gian nhất của quá trình khai phá tri thức, áp dụng các kỹ thuật khai phá phần lớn là các kỹ thuật của machine learning) để khai phá, trích chọn được các mẫu (pattern) thông tin, các mối liên hệ đặc biệt trong dữ liệu.

Bước 5: Đánh giá và biểu diễn tri thức (Knowledge representation & evaluation) Dùng các kỹ thuật hiển thị dữ liệu để trình bày các thông tin (tri thức) và mối liên hệ đặc biệt trong dữ liệu đã được khai thác ở bước trên biểu diễn dưới dạng gần gũi với người sử dụng như đồ thị, cây, bảng biểu, luật…Đồng thời bước này cũng đánh giá những 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 người dung để điều chỉnh và rút ra các tri thức cần thiết. Các tri thức nhận được cũng có thể được lưu và sử dụng lại.



Hình 1.1: Qúa trình phát hiện tri thức

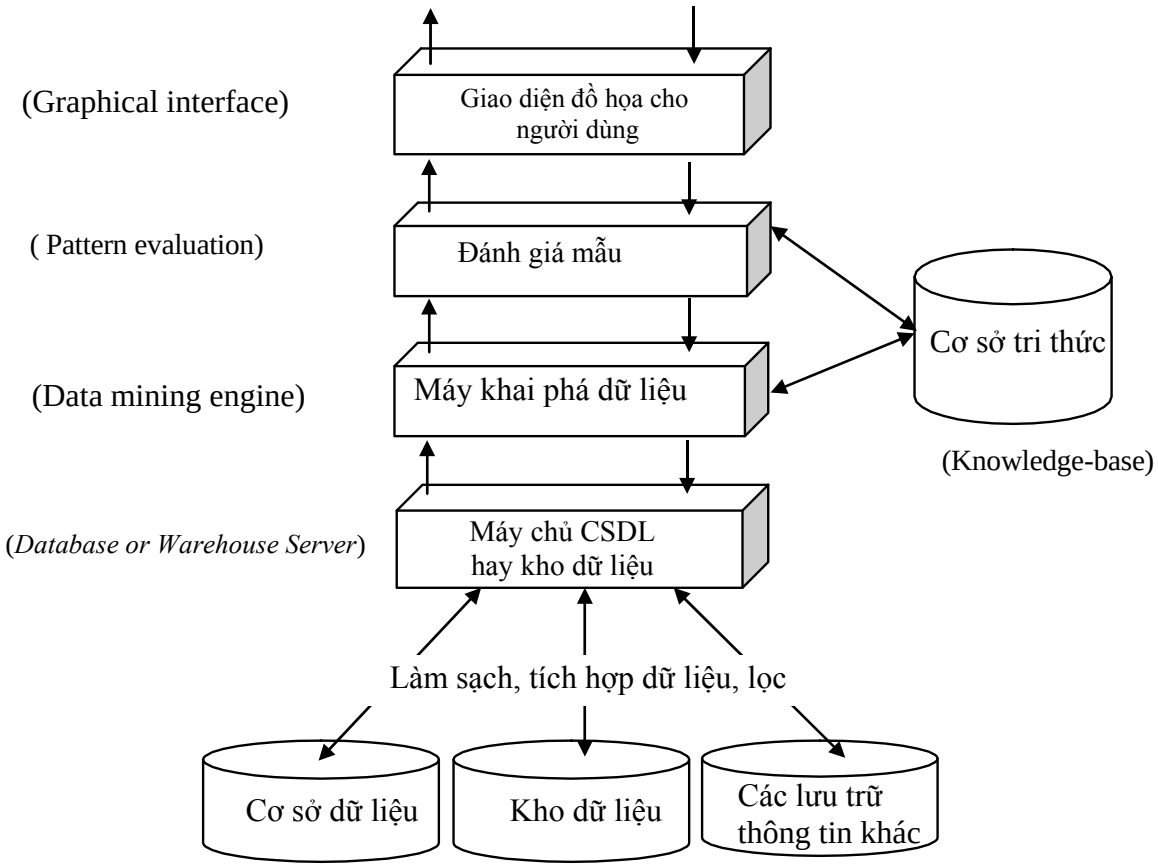
Việc KPDL có thể được tiến hành trên một lượng lớn dữ liệu có trong các CSDL (Cơ sở dữ liệu), các kho dữ liệu hoặc trong các loại lưu trữ thông tin khác.

Các mẫu đáng quan tâm có thể được đưa đến người dung hoặc được lưu trữ trong một số cơ sở tri thức.

1.2. Kiến trúc của một hệ thống khai phá dữ liệu

Kiến trúc của một hệ thống KPDL điển hình có thể có các thành phần như hình 1.2, [5], [9]

CSDL, kho dữ liệu hoặc các lưu trữ thông tin khác (**Databases, Data warehouse,..).** Đây là một hay một tập các CSDL, các kho dữ liệu, các trang tính hay các dạng lưu trữ thông tin khác. Các kỹ thuật làm sạch dữ liệu và tích hợp dữ liệu có thể được thực hiện trên những dữ liệu này.



Hình 1.2: Kiến trúc của một hệ thống khai phá dữ liệu

- *Máy chủ CSDL hay máy chủ kho dữ liệu (Database or warehouse server).* Máy chủ này có trách nhiệm lấy dữ liệu thích hợp dựa trên các yêu cầu khai phá của người dùng.

- *Cơ sở tri thức (Knowledge base).* Đây là nhiều tri thức được dùng để hướng dẫn việc tìm kiếm hay đánh giá độ quan trọng của các hình mẫu kết quả.

- *Máy KPDL (Data mining engine)* Một hệ thống KPDL cần phải có một tập các modun chức năng để thực hiện công việc như: đặc trưng hóa, kết hợp, phân lớp, phân cụm, phân tích sự tiến hóa.

- *Modun đánh giá mẫu (Pattern evaluation).* Bộ phận này tương tác với các modun KPDL để duyệt tìm các mẫu đáng được quan tâm. Nó có thể dung các ngưỡng về độ quan tâm để lọc mẫu đã khám phá được. Cũng có thể modun đánh giá mẫu được tích hợp vào modun khám phá, tùy theo sự cài đặt của phương pháp

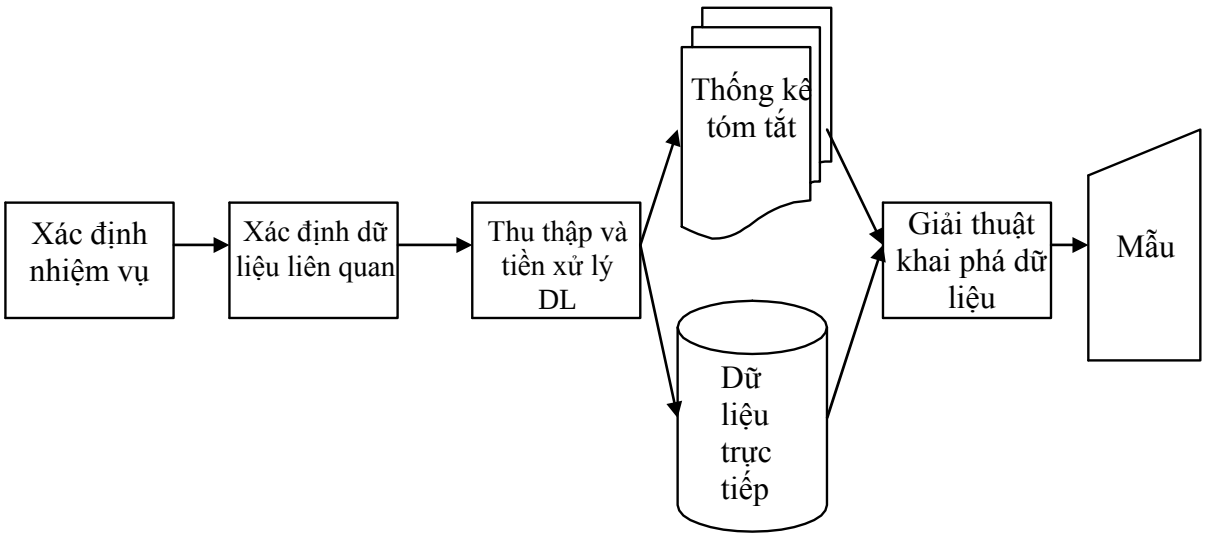
khai phá được dùng.

- *Giao diện người dung (Graphical user interface).* Bộ phận này cho phép người dùng giao tiếp với hệ thống KPDL. Ngoài ra bộ phận này còn cho phép người dung xem các lược đồ CSDL, lược đồ kho dữ liệu (hay các cấu trúc dữ liệu), các đánh giá mẫu và hiển thị các mẫu trong khuôn dạng khác nhau.

1.3 Các giai đoạn của quá trình khai phá dữ liệu

Các giải thuật khai phá dữ liệu thường được miêu tả như những chương trình hoạt động trực tiếp trên tệp dữ liệu. Với các phương pháp học máy và thống kê trước đây, thường thì bước đầu tiên là các giải thuật nạp toàn bộ tệp dữ liệu vào trong bộ nhớ. Khi chuyển sang các ứng dụng công nghiệp liên quan đến việc khai phá các kho dữ liệu, mô hình này không thể đáp ứng được. Không chỉ bởi vì nó không thể nạp hết dữ liệu vào trong bộ nhớ mà còn vì khó có thể chiết xuất dữ liệu ra các tệp đơn giản để phân tích được.

Quá trình khai phá dữ liệu được thể hiện bởi mô hình sau [3]:



Hình 1.3. Quá trình khai phá dữ liệu

+ Xác định nhiệm vụ: Xác định chính xác vấn đề cần giải quyết.

+ Xác định dữ liệu liên quan: Dùng để xây dựng giải pháp.

+ Thu thập các dữ liệu có liên quan và xử lý chúng thành dạng sao cho giải thuật khai phá dữ liệu có thể hiểu được. Ở đây có thể gặp một số vấn đề: dữ liệu phải được sao ra nhiều bản (nếu được chiết xuất vào các tệp), quản lý tập các tệp dữ

liệu, phải lặp đi lặp lại nhiều lần toàn bộ quá trình (nếu mô hình dữ liệu thay đổi v.v...).

+ Chọn thuật toán khai phá dữ liệu thích hợp và thực hiện việc khai phá dữ liệu: nhằm tìm được các mẫu (pattern) có ý nghĩa dưới dạng biểu diễn tương ứng với các ý nghĩa đó.

1.4. Một số kỹ thuật khai phá dữ liệu

Mục đích của khai phá dữ liệu là chiết xuất ra các tri thức có lợi cho kinh doanh hay cho nghiên cứu khoa học… Do đó, ta có thể xem mục đích của khai phá dữ liệu sẽ là mô tả các sự kiện và dự đoán. Các mẫu khai phá dữ liệu phát hiện được nhằm vào mục đích này. Dự đoán liên quan đến việc sử dụng các biến hoặc các đối tượng (bản ghi) trong CSDL để chiết xuất ra các mẫu, dự đoán được những giá trị chưa biết hoặc những giá trị tương lai của các biến đáng quan tâm. Mô tả tập trung vào việc tìm kiếm các mẫu mô tả dữ liệu mà con người có thể hiểu được.

Một số kỹ thuật phổ biến thường được sử dụng để KPDL hiện nay là :

Phân lớp dữ liệu

Mục tiêu của phân lớp dữ liệu 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. Mô hình được sử dụng để dự đoán nhãn lớp khi mà độ chính xác của mô hình chấp nhận được.

Phân nhóm dữ liệu

Phân nhóm là kỹ thuật khai phá dữ liệu tương tự như phân lớp dữ liệu. Tuy nhiên, sự phân nhóm dữ liệu là quá trình học không được giám sát, là quá trình nhóm những đối tượng vào trong những lớp tương đương, đến những đối tượng trong một nhóm là tương đương nhau, chúng phải khác với những đối tượng trong những nhóm khác. Trong phân lớp dữ liệu, một bản ghi thuộc về lớp nào là phải xác định trước, trong khi phân nhóm không xác định trước Trong phân nhóm, những đối tượng được nhóm lại cùng nhau dựa vào sự giống nhau của chúng. Sự giống nhau giữa những đối tượng được xác định

bởi những chức năng giống nhau. Thông thường những sự giống nhau về định lượng như khoảng cách hoặc độ đo khác được xác định bởi những chuyên gia trong lĩnh vực của mình.

Đa số các ứng dụng phân nhóm được sử dụng trong sự phân chia thị trường. Với sự phân nhóm khách hàng vào trong từng nhóm, những doanh nghiệp có thể cung cấp những dịch vụ khác nhau tới nhóm khách hàng một cách thuận lợi. Ví dụ: dựa vào chi tiêu, số tiền trong tài khoản và việc rút tiền của khách hàng, một ngân hàng có thể xếp những khách hàng vào những nhóm khác nhau. Với mỗi nhóm, ngân hàng có thể cho vay những khoản tiền tương ứng cho việc mua nhà, mua xe, … Trong trường hợp này ngân hàng có thể cung cấp những dịch vụ tốt hơn, và cũng chắc chắn rằng tất cả các khoản tiền cho vay đều có thể thu hồi được. Ta có thể tham khảo một khảo sát toàn diện về kỹ thuật và thuật toán phân nhóm trong.

Khai phá luật kết hợp

Mục tiêu của phương pháp này là phát hiện và đưa ra 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 thường xuyên. Một tập mục thường xuyên được xác định thông qua việc tính độ hỗ trợ và thoả mãn độ hỗ trợ cực tiểu.

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

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.

Phát hiện sự thay đổi và độ lệch (change and deviation dectection): Nhiệm

vụ này tập trung vào khám phá những thay đổi có ý nghĩa trong dữ liệu dựa vào các giá trị chuẩn hay độ đo đã biết trước, phát hiện độ lệch đáng kể giữa

nội dung của tập con dữ liệu và nội dung mong đợi. Hai mô hình độ lệch thường dùng là lệch theo thời gian và lệch theo nhóm. Độ lệch theo thời gian là sự thay đổi có nghĩa của dữ liệu theo thời gian. Độ lệch theo nhóm là sự khác nhau giữa dữ liệu trong hai tập con dữ liệu, tính cả trường hợp tập con của đối tượng này thuộc tập con kia, nghĩa là xác định dữ liệu trong một nhóm con của đối tượng có khác nhau đáng kể so với toàn bộ đối tượng.

1.5. Các cơ sở dữ liệu phục vụ cho khai phá dữ liệu.

Dựa vào những kiểu dữ liệu mà kỹ thuật khai phá áp dụng, có thể chia dữ liệu thành các loại khác nhau.

Cơ sở dữ liệu quan hệ

Đến nay, hầu hết dữ liệu được lưu giữ dưới dạng cơ sở dữ liệu quan hệ. Cơ sở dữ liệu quan hệ là một nguồn tài nguyên lớn nhất chứa những đối tượng mà chúng ta cần khai phá. Cơ sở dữ liệu quan hệ có cấu trúc cao, dữ liệu được mô tả bởi một tập những thuộc tính và lưu trong những bảng. Khai phá dữ liệu trên cơ sở dữ liệu quan hệ chủ yếu tập trung khai phá mẫu. Ví dụ, trong cơ sở dữ liệu của một ngân hàng, ta có thể tìm được những khách hàng có mức chi tiêu cao, ta có thể phân loại những khách hàng này dựa vào quá trình chi tiêu của họ. Cũng với việc phân tích những mục chi tiêu của khách hàng, chúng ta có thể cung cấp một số thông tin của khách hàng đến những doanh nghiệp khác. Giả sử rằng một khách hàng chi mỗi tháng 500 đô la cho thời trang, nếu được phép, ngân hàng có thể cung cấp thông tin về khách hàng này cho những cửa hàng thời trang.

Cơ sở dữ liệu giao tác

Cơ sở dữ liệu giao tác là tập hợp những bản ghi giao dịch, trong đa số các trường hợp chúng là những bản ghi các dữ liệu hoạt động của doanh nghiệp, tổ chức. Với tính phổ biến của máy tính và thương mại điện tử, ngày nay có rất nhiều cơ sở dữ liệu giao tác. Khai phá dữ liệu trên cơ sở dữ liệu giao tác tập trung vào khai phá lật kết hợp, tìm mối tương quan giữa những mục dữ liệu của bản ghi giao dịch. Nghiên cứu sâu về cơ sở dữ liệu giao tác được mô tả chi tiết ở phần sau.

Cơ sở dữ liệu không gian

Cơ sở dữ liệu không gian bao gồm hai phần: Phần thứ nhất là dữ liệu quan hệ hay giao tác, phần thứ hai là thông tin định vị hoặc thông tin địa lý. Những luật kết hợp trên cơ sở dữ liệu không gian mô tả mối quan hệ giữa các đặc trưng trong cơ sở dữ liệu không gian. Dạng của luật kết hợp không gian có dạng X => Y, với X, Y là tập hợp những vị từ không gian. Những thuật toán khai phá luật kết hợp không gian tương tự như khai phá luật kết hợp nhưng thêm những vị từ về không gian.

Cơ sở dữ liệu có yếu tố thời gian

Giống như cơ sở dữ liệu không gian, cơ sở dữ liệu có yếu tố thời gian bao gồm hai phần: Phần thứ nhất là dữ liệu quan hệ hay giao tác, phần thứ hai là thông tin về thời gian xuất hiện dữ liệu ở phần thứ nhất. Những luật kết hợp có yếu tố thời gian có nhiều thông tin hơn những luật kết hợp cơ bản. Ví dụ, từ luật kết hợp cơ bản {Bia}⇒ {Thuốc lá}, với dữ liệu có yếu tố thời gian chúng ta có thể có nhiều luật: Độ hỗ trợ của luật {Bia}⇒ {Thuốc lá} là 20% từ 9 giờ đến 13 giờ, là 50% trong thời gian 19 giờ tới 22 giờ. Rõ ràng rằng, những người bán lẻ có thể xác định chiến lược để buôn bán tốt hơn.

Hầu hết nghiên cứu về lĩnh vực này ngày nay hình thành một hướng khai phá dữ liệu mới gọi là khai phá mẫu lặp liên tục, khai phá tập mục dữ liệu thường xuyên trong cơ sở dữ liệu thời gian.

Cơ sở dữ liệu đa phương tiện

Số lượng trang web đang bùng nổ trên thế giới, web có mặt ở khắp mọi nơi, duyệt web đã là nhu cầu của mọi tầng lớp trong xã hội. Thông tin trên web đang phát triển với tốc độ rất cao, khai phá thông tin trên web (web mining) đã trở thành một lĩnh vực nghiên cứu chính của khai phá dữ liệu, được các nhà nghiên cứu đặc biệt quan tâm.

Khai phá dữ liệu web thông thường được chia thành ba phạm trù chính: Khai phá cách dùng web (web usage mining), khai phá c ấu trúc web (web structure mining) và khai phá nội dung web (web content mining).