**ĐẠI HỌC THÁI NGUYÊN**
**KHOA CÔNG NGHỆ THÔNG TIN**

**LÊ THU HÀ**

# PHƯƠNG PHÁP LUẬN KẾT HỢP VÀ ỨNG DỤNG

**Luận văn thạc sỹ : Khoa học máy tính**

**Thái Nguyên - 2009**

**ĐẠI HỌC THÁI NGUYÊN**
**KHOA CÔNG NGHỆ THÔNG TIN**

**LÊ THU HÀ**

# PHƯƠNG PHÁP LUẬN KẾT HỢP VÀ ỨNG DỤNG

Chuyên ngành: **: Khoa học máy tính**

Mã số: **60 48 01**

**Luận văn Thạc sỹ Khoa học máy tính**

**NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS VŨ ĐỨC THI**

**Thái Nguyên - 2009**

# MỤC LỤC

# MỞ ĐẦU

Trong những năm gần đây, sự phát triển mạnh mẽ của công nghệ thông tin đã làm cho khả năng thu thập và lưu trữ thông tin của các hệ thống thông tin tăng nhanh một cách nhanh chóng. Bên cạnh đó, việc tin học hóa một cách ồ ạt và nhanh chóng các hoạt động sản xuất, kinh doanh cũng như nhiều lĩnh vực hoạt động khác đã tạo ra cho chúng ta một lượng dữ liệu cần lưu trữ khổng lồ. Hàng triệu cơ sở dữ liệu đã được sử dụng trong các hoạt động sản xuất, kinh doanh, quản lý..., trong đó có nhiều cơ sở dữ liệu cực lớn cỡ Gigabyte, thậm chí là Terabyte.

Sự bùng nổ này đã dẫn tới một yêu cầu cấp thiết là cần có những kỹ thuật và công cụ mới để tự động chuyển đổi lượng dữ liệu khổng lồ kia thành các tri thức có ích. Từ đó, các kỹ thuật khai phá dữ liệu đã trở thành một lĩnh vực thời sự của nền CNTT thế giới hiện nay nói chung và Việt Nam nói riêng.

Khai phá dữ liệu đang được áp dụng một cách rộng rãi trong nhiều lĩnh vực kinh doanh và đời sống khác nhau: marketing, tài chính, ngân hàng và bảo hiểm, khoa học, y tế, an ninh, internet… Rất nhiều tổ chức và công ty lớn trên thế giới đã áp dụng kỹ thuật khai phá dữ liệu vào các hoạt động sản xuất kinh doanh của mình và thu được những lợi ích to lớn.

Mục đích nghiên cứu của đề tài là tìm hiểu về các kỹ thuật khai phá dữ liệu; các vấn đề liên quan đến khai phá luật kết hợp nhằm phát hiện và đưa ra các mối liên hệ giữa các giá trị dữ liệu trong CSDL và áp dụng chúng vào bài toán quản lý trang thiết bị đồ dùng của trường THPT Chu Văn An – Tỉnh Thái Nguyên.

*Mục tiêu nghiên cứu của đề tài:*

- Tổng kết các kiến thức cơ bản nhất liên quan đến phát hiện luật kết hợp và tìm kiếm tri thức từ dữ liệu.

- Dựa trên lý thuyết đã tổng kết được, đi sâu vào tìm hiểu, nghiên cứu phương pháp luật kết hợp và làm một chương trình thử nghiệm dựa trên thuật toán Apriori.

*Ý nghĩa khoa học của đề tài:*

- Đây là phương pháp được nhiều nhà khoa học nghiên cứu và đã có đóng góp trong thực tiễn.
- Có thể coi đề tài là một tài liệu tham khảo khá đầy đủ, rõ ràng về các kiến thức cơ bản trong phương pháp phát hiện luật kết hợp.

*Phương pháp nghiên cứu:*

- Lập kế hoạch, lên qui trình, tiến độ thực hiện.
- Tham khảo nhiều tài liệu có liên quan, tham khảo ý kiến các chuyên gia trong lĩnh vực nghiên cứu.

*Phạm vi nghiên cứu:*

Các kiến thức cơ bản nhất về phương pháp phát hiện luật kết hợp trên cơ sở làm luận văn thạc sỹ.

*Các kết quả nghiên cứu đạt được:*

- Tổng kết các kiến thức cơ bản nhất của phương pháp khai phá luật kết hợp.
- Luận văn có thể trở thành một tài liệu tham khảo cho những người muốn tìm hiểu về khai phá dữ liệu và phương pháp khai phá luật kết hợp.
- Xây dựng một phần mềm thử nghiệm dựa trên thuật toán Apriori.

Luận văn bao gồm 4 chương, với các nội dung:

Chương 1: Trình bày tổng quan về khám phá tri thức và khai phá dữ liệu, trong đó có đề cập đến khái niệm tri thức, dữ liệu, quá trình khám phá tri thức, nhiệm vụ và các kỹ thuật khám phá tri thức.

Chương 2: Trình bày về luật kết hợp, trong đó trình bày về các khái niệm, định nghĩa, tính chất của luật kết hợp.

Chương 3: Trình bày một số kỹ thuật khai thác luật kết hợp.

Chương 4: Cài đặt chương trình tìm luật kết hợp, ứng dụng trong quản lý trang thiết bị, đồ dùng của trường THPT Chu Văn An – Tỉnh Thái Nguyên.

Luận văn này đã được hoàn thành trong khoảng thời gian không dài. Tuy nhiên, đã đạt được một số kết quả tốt, tôi đang nghiên cứu để hoàn thiện và đưa chương trình trong luận văn vào ứng dụng thực tiễn quản lý trang thiết bị của trường THPT Chu Văn An – Tỉnh Thái Nguyên, rất mong nhận được sự góp ý của các thầy cô, đồng nghiệp và bạn bè để luận văn và chương trình được hoàn thiện hơn.

# Chương 1
# TỔNG QUAN VỀ KHÁM PHÁ TRI THỨC VÀ KHAI PHÁ DỮ LIỆU

## 1.1. Phát hiện tri thức và khai phá dữ liệu

Trong thời đại bùng nổ công nghệ thông tin, các công nghệ lưu trữ dữ liệu ngày càng phát triển tạo điều kiện cho các đơn vị thu thập dữ liệu tốt hơn. Đặc biệt trong lĩnh vực kinh doanh, các doanh nghiệp đã nhận thức được tầm quan trọng của việc nắm bắt và xử lý thông tin, nhằm giúp các chủ doanh nghiệp trong việc vạch ra các chiến lược kinh doanh kịp thời mang lại những lợi nhuận to lớn cho doanh nghiệp của mình. Tất cả lí do đó khiến cho các cơ quan, đơn vị và các doanh nghiệp đã tạo ra một lượng dữ liệu khổng lồ cỡ Gigabyte thậm chí là Terabyte cho riêng mình.

Khi lưu trữ các dữ liệu khổng lồ như vậy thì chúng ta thấy rằng chắc chắn chúng phải chứa những giá trị nhất định nào đó. Tuy nhiên, theo thống kê thì chỉ có một lượng nhỏ của những dữ liệu này (khoảng từ 5% đến 10%) là luôn được phân tích, số còn lại họ không biết sẽ phải làm gì hoặc có thể làm gì với chúng nhưng họ vẫn tiếp tục thu thập rất tốn kém với ý nghĩ lo sợ rằng sẽ có cái gì đó quan trọng đã bị bỏ qua sau này có lúc cần đến nó. Mặt khác, trong môi trường cạnh tranh, người ta ngày càng cần có nhiều thông tin với tốc độ nhanh để trợ giúp việc ra quyết định và ngày càng có nhiều câu hỏi mang tính chất định tính cần phải trả lời dựa trên một khối lượng dữ liệu khổng lồ đã có. Với những lý do như vậy, các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được thực tế đã làm phát triển một khuynh hướng kỹ thuật mới đó là Kỹ thuật phát hiện tri thức và khai phá dữ liệu (KDD - Knowledge Discovery and Data Mining).

Thông thường chúng ta coi dữ liệu như một dãy các bit, hoặc các số và các ký hiệu, hoặc các "đối tượng" với một ý nghĩa nào đó khi được gửi cho một chương trình dưới một dạng nhất định. Chúng ta sử dụng các bit để đo

lường các thông tin và xem nó như là các dữ liệu đã được lọc bỏ các dư thừa, được rút gọn tới mức tối thiểu để đặc trưng một cách cơ bản cho dữ liệu. Chúng ta có thể xem tri thức như là các thông tin tích hợp, bao gồm các sự kiện và các mối quan hệ giữa chúng. Các mối quan hệ này có thể được hiểu ra, có thể được phát hiện, hoặc có thể được học. Nói cách khác, tri thức có thể được coi là dữ liệu có độ trừu tượng và tổ chức cao.

Phát hiện tri thức trong các cơ sở dữ liệu 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. Còn khai thác dữ liệu là một bước trong qui trình phát hiện tri thức gồm có các thuật toán khai thác dữ liệu chuyên dùng dưới một số qui định về hiệu quả tính toán chấp nhận được để tìm ra các mẫu hoặc các mô hình trong dữ liệu. Nói một cách khác, mục đích của phát hiện tri thức và khai phá dữ liệu chính là tìm ra các mẫu và/hoặc các mô hình đang tồn tại trong các cơ sở dữ liệu nhưng vẫn còn bị che khuất bởi hàng "núi" dữ liệu.

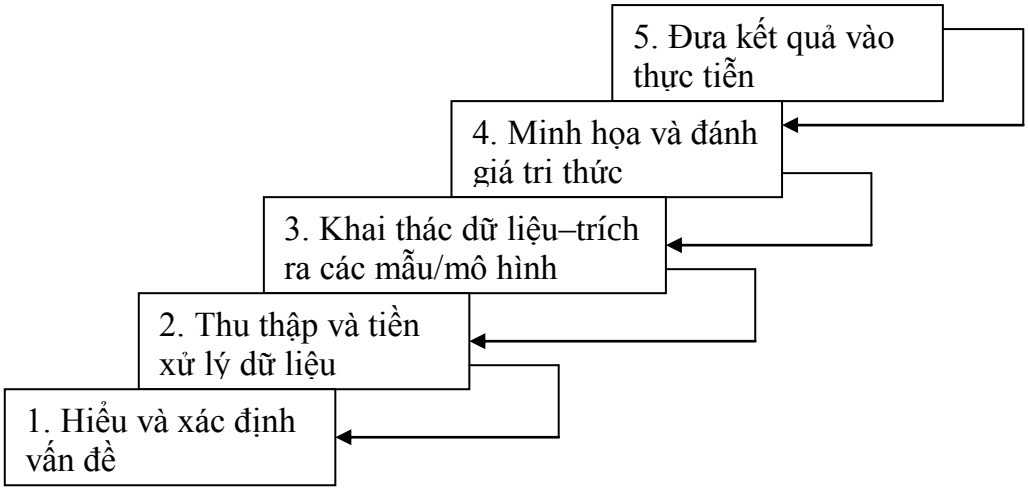
Nhiều người coi khai phá dữ liệu và khám phá tri thức trong cơ sở dữ liệu là như nhau. Tuy nhiên trên thực tế, khai phá dữ liệu chỉ là một bước thiết yếu trong quá trình phát hiện tri thức trong cơ sở dữ liệu.

### 1.2. Quá trình phát hiện tri thức từ cơ sở dữ liệu

Quá trình phát hiện tri thức có thể chia thành các bước như sau:

- Làm sạch dữ liệu (Data cleaning): Loại bỏ dữ liệu nhiễu hoặc dữ liệu không thích hợp.
- Tích hợp dữ liệu (Data integration): Tích hợp dữ liệu từ các nguồn khác nhau.
- Chọn dữ liệu (Data Selection): Chọn những dữ liệu liên quan trực tiếp đến nhiệm vụ.
- Chuyển đổi dữ liệu (Data Transformation): Chuyển dữ liệu về những dạng phù hợp cho việc khai phá.

- Khai phá dữ liệu (Data mining): Các kỹ thuật được áp dụng để trích xuất thông tin có ích hoặc các mẫu điển hình trong dữ liệu.
- Đánh giá mẫu (Pattern evaluation): Đánh giá mẫu hoặc tri thức đã thu được.
- Trình diễn dữ liệu (Knowledge Presentation): Biểu diễn những tri thức khai phá được cho người sử dụng.




***Hình 1.1. Quá trình khám phá tri thức từ cơ sở dữ liệu***

Hình 1.1 mô tả 5 giai đoạn trong quá trình khám phá tri thức từ cơ sở dữ liệu. Mặc dù có 5 giai đoạn như trên xong quá trình khám phá tri thức từ cơ sở dữ liệu là một quá trình tương tác và lặp di lặp lại theo chu trình liên tục kiểu xoáy trôn ốc, trong đó lần lặp sau hoàn chỉnh hơn lần lặp trước. Ngoài ra, giai đoạn sau lại dựa trên kết quả thu được của giai đoạn trước theo kiểu thác nước. Đây là một quá trình biện chứng mang tính chất khoa học của lĩnh vực phát hiện tri thức và là phương pháp luận trong việc xây dựng các hệ thống phát hiện tri thức.

### 1.2.1. Xác định vấn đề

Đây là một quá trình mang tính định tính với mục đích xác định được lĩnh vực yêu cầu phát hiện tri thức và xây dựng bài toán tổng kết. Trong thực tế,