ĐẠI HỌC THÁI NGUYÊN
KHOA CÔNG NGHỆ THÔNG TIN

AN HỒNG SƠN

NGHIÊN CỨU MỘT SỐ PHƯƠNG PHÁP PHÂN CỤM MỜ VÀ ỨNG DỤNG

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH
MÃ SỐ: 60 48 01

LUẬN VĂN THẠC SĨ KHOA HỌC

HƯỚNG DẪN KHOA HỌC: PGS.TS NGÔ QUỐC TẠO

THÁI NGUYÊN - 2008

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

CNTT	Công nghệ thông tin
CSDL	Cơ sở dữ liệu
CEF	Computational Energy Function
DL	Dữ liệu
FBACN	Fuzzy Bi-directional Associative Clustering Network *(Mạng Nơron đa khớp phục vụ cho phân cụm mờ)*
FCM	Fuzzy C-Means
HMT	Hàm mục tiêu
KPDL	Khai phá dữ liệu
KPTT	Khám phá tri thức
LKM	Liên kết mạng
MH	Mô hình
NDA	Nhận dạng ảnh
NN	Neural Network
PCM	Phân cụm mờ
PCDL	Phân cụm dữ liệu
TLTK	Tài liệu tham khảo
TT	Thuật toán
XLA	Xử lý ảnh

DANH MỤC CÁC HÌNH MINH HOẠ

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

1.1. Giới thiệu chung về khám phá tri thức và khai phá dữ liệu

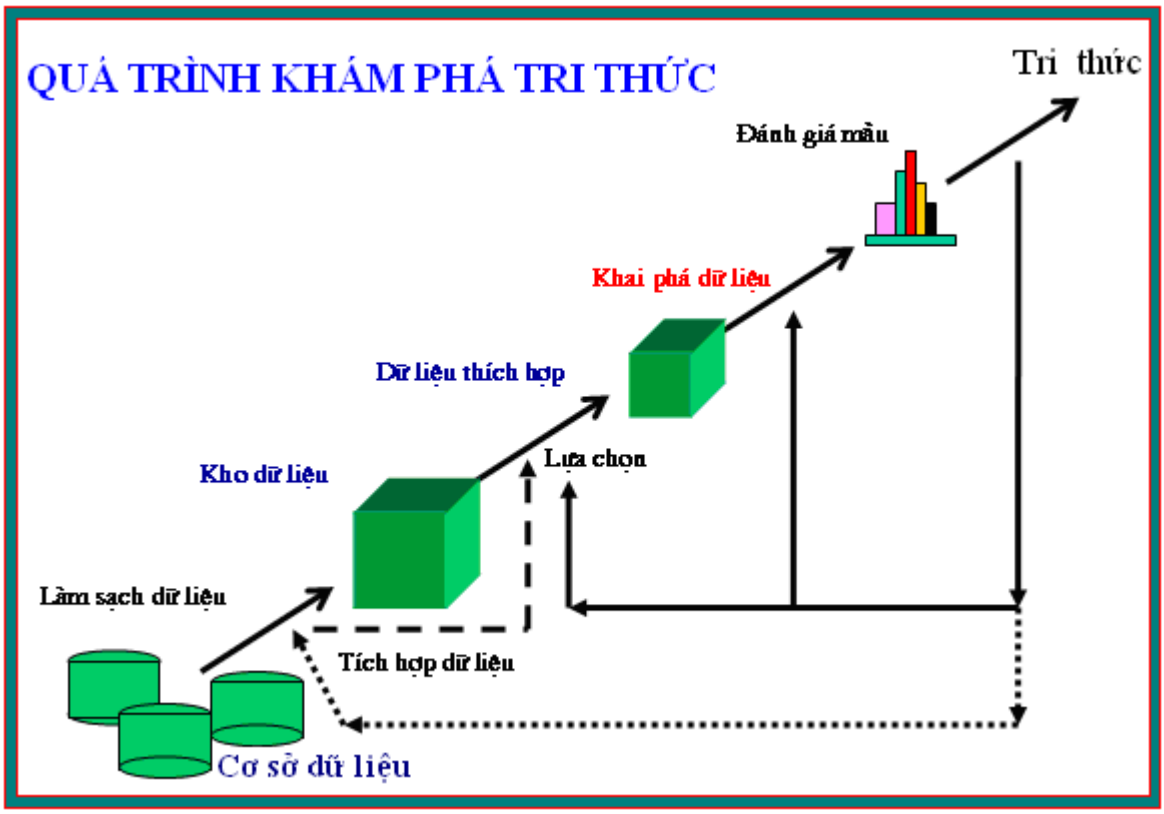
Nếu cho rằng, điện tử và truyền thông chính là bản chất của khoa học điện tử, thì dữ liệu, thông tin, và tri thức hiện đang là tiêu điểm của một lĩnh vực mới để nghiên cứu và ứng dụng, đó là khám phá tri thức và khai phá dữ liệu.

Thông thường, chúng ta coi *dữ liệu* như là một chuỗi các bits, hoặc các số và các ký hiệu hay là các "đối tượng" với một ý nghĩa nào đó khi được gửi cho một chương trình dưới một dạng nhất định. Các bits thường được sử dụng để đo *thông tin*, và xem nó như là dữ liệu đã được loại bỏ phần tử thừa, lặp lại, và rút gọn tới mức tối thiểu để đặc trưng một cách cơ bản cho dữ liệu. *Tri thức* được xem như là các thông tin tích hợp, bao gồm các sự kiện và mối quan hệ giữa chúng, đã được nhận thức, khám phá, hoặc nghiên cứu. Nói cách khác, tri thức có thể được coi là dữ liệu ở mức độ cao của sự trừu tượng và tổng quát.

Khám phá tri thức hay phát hiện tri thức trong CSDL là một quy trình nhận biết các mẫu hoặc các mô hình trong dữ liệu với các tính năng: Phân tích, tổng hợp, hợp thức, khả ích và có thể hiểu được.

Khai phá dữ liệu là một bước trong quá trình khám phá tri thức, gồm 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 cách khác, mục tiêu của Khai phá dữ liệu là tìm kiếm các mẫu hoặc mô hình tồn tại trong CSDL nhưng ẩn trong khối lượng lớn dữ liệu.

1.2. Quá trình khám phá tri thức



Hình 1.1: Quá trình KPTT

Bao gồm các bước sau:

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

Tích hợp dữ liệu (Data Intergation): Dữ liệu của nhiều nguồn có thể được tổ hợp lại.

Lựa chọn dữ liệu (Data Selection): Lựa chọn những dữ liệu phù hợp với nhiệm vụ phân tích trích rút từ cơ sở dữ liệu.

Chuyển đổi dữ liệu (Data Transformation): Dữ liệu được chuyển đổi hay được hợp nhất về dạng thích hợp cho việc khai phá.

Khai phá dữ liệu (Data Mining): Đây là một tiến trình cốt yếu trong đó các phương pháp thông minh được áp dụng nhằm trích rút ra mẫu dữ liệu.

Đánh giá mẫu (Pattern Evaluation): Dựa trên một độ đo nào đó xác định lợi ích thực sự, độ quan trọng của các mẫu biểu diễn tri thức.

Biểu diễn tri thức (Knowledge Presentation): Ở giai đoạn này các kỹ thuật biểu diễn và hiển thị được sử dụng để đưa tri thức lấy ra cho người dùng.

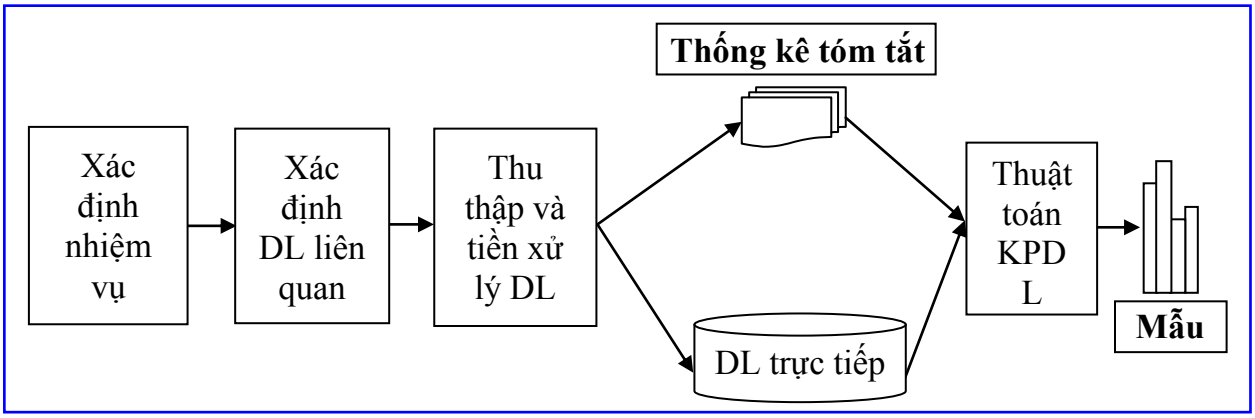
1.3. Quá trình khai phá dữ liệu

KPDL là một giai đoạn quan trọng trong quá trình KPTT. Về bản chất, nó là giai đoạn duy nhất tìm ra được thông tin mới, thông tin tiềm ẩn có trong CSDL chủ yếu phục vụ cho mô tả và dự đoán.

Mô tả dữ liệu là tổng kết hoặc diễn tả những đặc điểm chung của những thuộc tính dữ liệu trong kho dữ liệu mà con người có thể hiểu được.

Dự đoán là dựa trên những dữ liệu hiện thời để dự đoán những quy luật được phát hiện từ các mối liên hệ giữa các thuộc tính của dữ liệu trên cơ sở đó chiết xuất ra các mẫu, dự đoán được những giá trị chưa biết hoặc những giá trị tương lai của các biến quan tâm.

Quá trình KPDL bao gồm các bước chính được thể hiện như Hình 1.2 sau:



Hình 1.2: Quá trình KPDL

- *Xác định nhiệm vụ:* Xác định chính xác các vấn đề cần giải quyết.
- *Xác định các dữ liệu liên quan:* Dùng để xây dựng giải pháp.
- *Thu thập và tiền xử lý dữ liệu:* Thu thập các dữ liệu liên quan và tiền xử lý chúng sao cho thuật toán KPDL có thể hiểu được. Đây là một quá trình rất khó khăn, có thể gặp phải rất nhiều các vướng mắc như: dữ liệu phải được sao ra nhiều bản (nếu được chiết xuất vào các tệp), quản lý tập các dữ liệu, phải lặp đi lặp lại nhiều lần toàn bộ quá trình (nếu mô hình dữ liệu thay đổi), v.v..
- *Thuật toán khai phá dữ liệu:* Lựa chọn thuật toán KPDL và thực hiện việc PKDL để tìm được các mẫu có ý nghĩa, các mẫu này được biểu diễn dưới dạng luật kết hợp, cây quyết định... tương ứng với ý nghĩa của nó.

1.4. Các phương pháp khai phá dữ liệu

Với hai mục đích khai phá dư liệu là Mô tả và Dự đoán, người ta thường sử dụng các phương pháp sau cho khai phá dữ liệu:

- Luật kết hợp *(association rules)*
- Phân lớp *(Classfication)*
- Hồi qui *(Regression)*
- Trực quan hóa *(Visualiztion)*