Bài 1:
Tổng quan về khai phá dữ liệu

PGS. TS. Đỗ Phúc
Khoa Hệ thống thông tin
Trường Đại học Công nghệ Thông tin, ĐHQG-HCM

Khai phá dữ liệu

- **Có sẵn khối dữ liệu lớn:**
  - **Các CSDL khổng lồ**
  - **Dữ liệu từ Internet**

Khai phá dữ liệu là gì ?

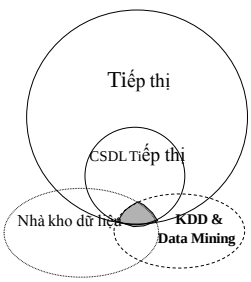
- **Rút trích thông tin hữu ích, chưa biết, tiềm ẩn trong khối dữ liệu lớn**
- **Phân tích dữ liệu bán tự động**
- **Giải thích dữ liệu trên các tập dữ liệu lớn .**

Khai phá dữ liệu là gì ?

- **Thuật ngữ:**
  - **Khai phá dữ liệu - Data mining**
    - **KPDL là một bước của tiến trình KDD**
  - **Knowledge discovery in databases (KDD)**
    - **Thuật ngữ tổng quát gồm các bước như tiền xử lý, KPDL, hậu xử lý .**

Khai phá dữ liệu có ích lợi gì ?

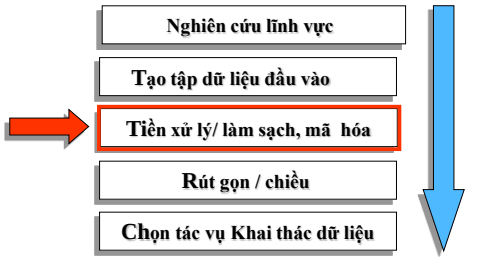
Cung cấp tri thức hỗ trợ ra quyết định
Dự báo
Khái quát dữ liệu



Các ứng dụng tiềm năng

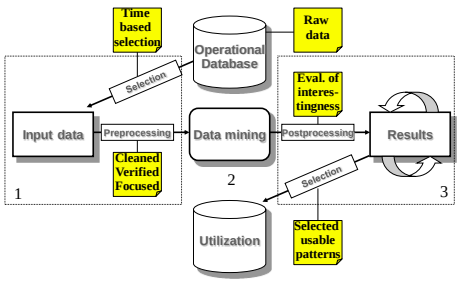
- **Phân tích dữ liệu, hỗ trợ ra quyết định**
  - Phân tích và quản lý thị trường
  - Quản lý và phân tích rủi ro
  - Quản lý và phân tích các sai hỏng
- **Các ứng dụng khác:**
  - Khai thác Web
  - Khai thác văn bản (text mining)
  - etc.

Tiến trình khai phá dữ liệu(1)

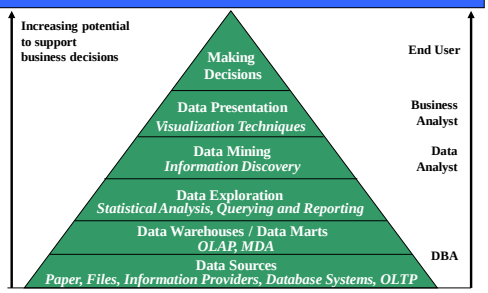


Tiến trình khai phá dữ liệu(2)

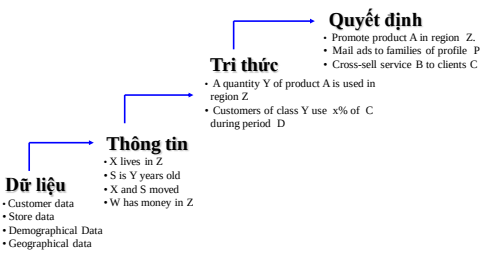
Tiến trình KDD tiêu biểu



Khai phá dữ liệu



Từ dữ liệu đến quyết định



Các quan niệm về KPDL

- **Các tiếp cận tổng quan:**
  - KPDL mô tả :
    - Cho biết điều gì là hữu ích có thể tìm thấy được trong dữ liệu
    - Giải thích dữ liệu đó
  - KPDL dự báo:
    - Dựa trên dữ liệu quá khứ, dự báo tương lai
    - Xu thế phát triển!

Các quan niệm về KTDL

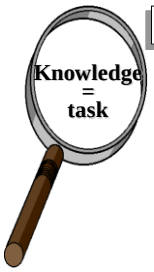
- **Quan niệm dựa trên …**
  - CSDL để khai thác
  - Tri thức được khám phá
  - Các kỹ thuật được sử dụng
  - Các ứng dụng

Các quan niệm về KPDL

CSDL cần khai thác

- Quan hệ
- Giao tác
- Hướng đối tượng
- Hướng đối tượng, quan hệ
- Active
- Không gian
- Thời gian
- Text, XML
- Multi-media
- Heterogeneous
- Legacy
- Inductive
- WWW
- etc.

Các quan niệm về KPDL



Tác vụ khai thác

- Đặc trưng
- Phân biệt
- Kết hợp
- Phân lớp
- Gom cụm
- Xu thế
- Phân tích độ lệch
- Phân tích hiếm etc.

Các quan niệm KPDL

Các kỹ thuật đã sử dụng

- **CSDL**
- **Nhà kho dữ liệu (OLAP)**
- **Máy học**
- **Thống kê**
- **Trực quan hóa**
- **Mạng nơron và thuật giải GA**
- **. . . .**

Các quan niệm về KPDL

Các ứng dụng

- Bán lẻ, siêu thị
- Ngân hàng
- Khai thác gen
- Phân tích cổ phiếu
- KTDL Web
- Phân tích dữ liệu

Kết luận

- **KPDL: tiến trình khám phá bán tự động các thông tin, mẫu có ích từ CSDL lớn**
- **Các bước của KDD**
  - Tiền xử lý
  - KTDL(data mining tasks)
  - Hậu xử lý
- **Các quan niệm, khía cạnh …**
  - CSDL (quan hệ, hướng đối tượng, không gian, WWW, …)
  - Tri thức (đặc trưng, gom cụm, kết hợp, …)
  - Kỹ thuật (máy học, thống kê, trực quan hóa, …)
  - Ứng dụng (bán lẻ, điện thoại, khai thác Web …)