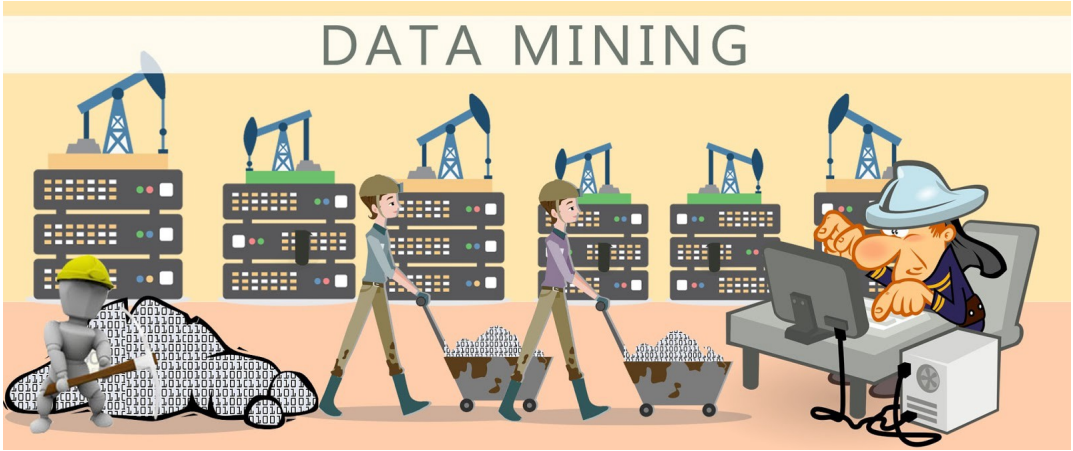
Xử lý ngôn ngữ tự nhiên

Chương 7: Data mining

Khoa CNTT, Đại học Kỹ thuật - Công nghệ Cần Thơ
Lưu hành nội bộ


DATA MINING

Data mining

- Data mining: khai phá /khoáng dữ liệu
- Tìm mẫu trong dữ liệu (find patterns in data)
- knowledge discovery in databases (KDD) - khám phá tri thức trong cơ sở dữ liệu

Khái niệm liên quan

- Machine learning
- Predictive analytics
- Big data
- Data science

Fayyad et. al. KDD process

- Data selection
- Data pre-processing
- Data transformation
- Data mining: the output is discovered patterns
- Data interpretation/evaluation

Han et. al. KDD process

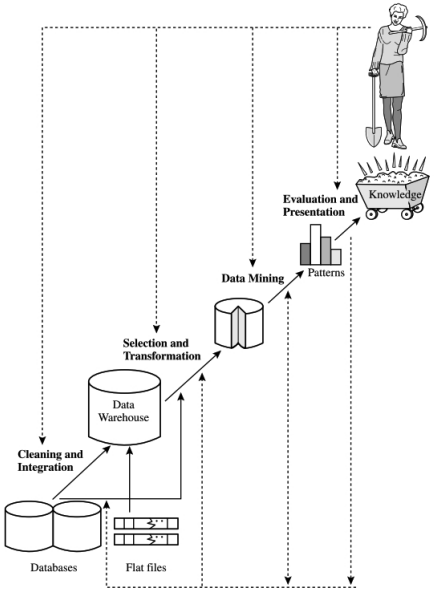
- Data cleaning
- Data integration
- Data selection
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge representation

The CRISP-DM process

- CRISP-DM: CRoss-Industry Standard Process for Data Mining
  1. Business understanding
  2. Data understanding
  3. Data preparation
  4. Modeling
  5. Evaluation
  6. Deployment

Techniques used in Data Mining

- Classification problems
- Clustering problems
- Regression problems
- Summarization problems
- Dependency modeling problems
- Change and deviation detection problems


Evaluation and Presentation
Knowledge
Data Mining
Patterns
Selection and Transformation
Data Warehouse
Cleaning and Integration
Databases
Flat files

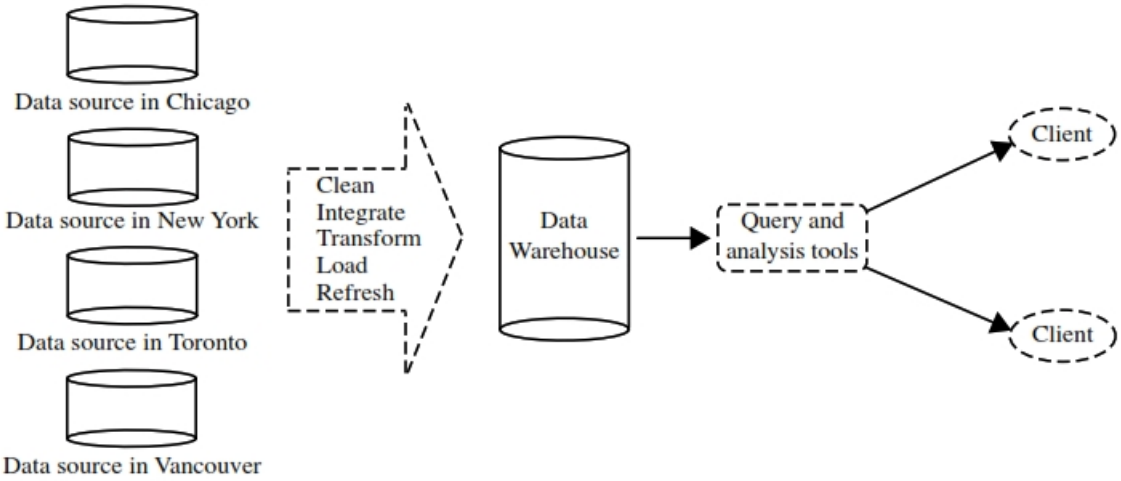
What kinds of data can be mined?

- data streams
- ordered/sequenced data
- graph or networked data
- spatial data
- text data
- multimedia data
- WWW

Database sources

- Relational databases
- Un-relational databases
- Data warehouses
- Transactional databases

Data warehouses



Descriptive function

- Class/Concept Description
- Mining of Frequent Patterns
- Mining of Associations
- Mining of Correlations
- Mining of Clusters

Class/Concept Description

- Data Characterization: summarizing data of class under study
- Data Discrimination: mapping or classification of a class with some predefined group or class

Mining of Frequent Patterns

- Frequent Item Set: a set of items that frequently appear together, for example, milk and bread.
- Frequent Subsequence: a sequence of patterns that occur frequently such as
  - purchasing a camera is followed by memory card
- Frequent Sub Structure - Substructure refers to different structural forms
  - graphs, trees, or lattices, which may be combined with item-sets or subsequences

Mining of clusters

- Hình thành các cluster dựa trên tối đa hóa sự tương đồng giữa các thành phần trong cluster
- Không biết trước bao nhiêu class label
- Mỗi cluster sau khi hình thành có thể xem như một class label

Evolution

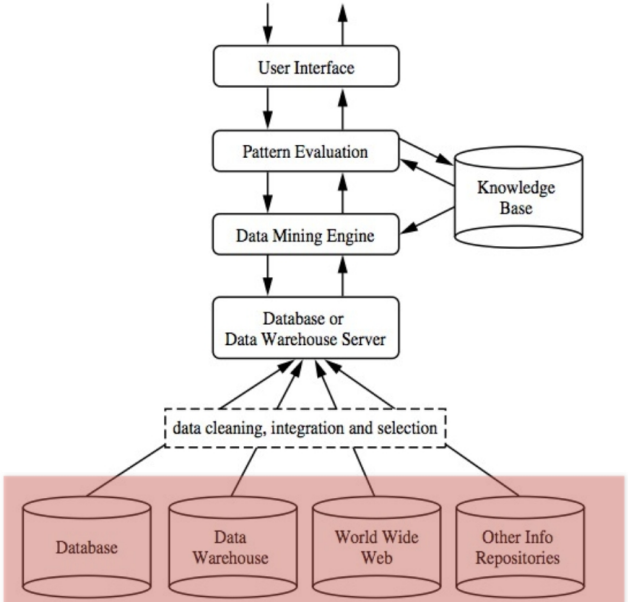
- Data mining có thể xem như tiến hóa (evolution) tự nhiên của công nghệ thông tin
- Kể từ những năm 1960, CSDL và CNTT đã tiến hóa từ file đến các hệ thống tận tiến và hệ thống CSDL
- 3 giai đoạn tiến hóa của CSDL

Information poorness

- Dư thừa dữ liệu (abundance of data)
- Đòi hỏi các công cụ phân tích dữ liệu
- Data rich, information poor
- Người dùng dựa vào các chuyên gia ngành để sàng lọc dữ liệu vào trở thành kiến thức
    - Quá trình dễ xảy ra thành kiến, lỗi, tốn thời gian, tốn chi phí

Data mining architecture

- User interface
- Pattern evaluation
- Data mining engine
- Knowledge base
- Database or data warehouse server
- Data sources


User Interface
Pattern Evaluation
Knowledge Base
Data Mining Engine
Database or Data Warehouse Server
data cleaning, integration and selection
Database
Data Warehouse
World Wide Web
Other Info Repositories

What kind of patterns can be mined?

- Data mining tasks can be classified into two categories:
  - Descriptive mining: tính chất chung của dữ liệu
  - Predictive mining: suy luận trên dữ liệu đang có, thực hiện dự đoán trên dữ liệu mới

Association analysis

- age(X, '20...29') $\wedge$ income(X, '200K...300K') $\Longrightarrow$ buys(X, 'milk tea')

  [support = 2%, confidence = 60%]

- Support
  - Khi qua sát tần suất (frequent) của các items trong dữ liệu, ta có tỉ lệ % xuất hiện cùng nhau
- Confidence
  - Trong các lần xuất hiện của các items, thì tỉ lệ % xảy ra vế imply

(a)

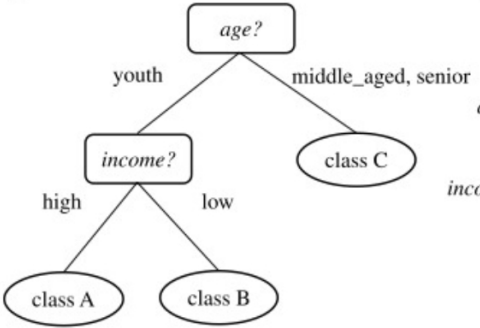
age(X, "youth") AND income(X, "high") ⟶ class(X, "A")

age(X, "youth") AND income(X, "low") ⟶ class(X, "B")

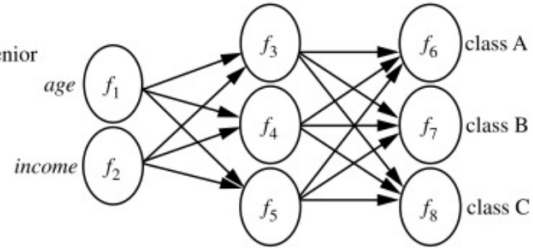
age(X, "middle_aged") ⟶ class(X, "C")

age(X, "senior") ⟶ class(X, "C")

(b)



(c)



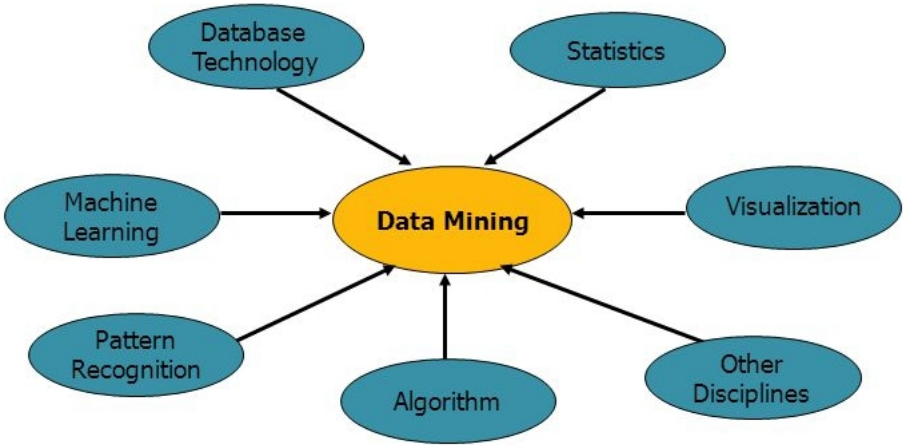
Outliers analysis

- Đối tượng dữ liệu (data objects) không tuân theo đặc tính, mô hình của dữ liệu
  - noise, exceptions
- Outliers được phát hiện bằng kiểm tra thống kê, sử dụng phép đo đối tượng đến các cluster

Interesting patterns

- Chỉ một phần nhỏ dữ liệu có tiềm năng là sự quan tâm của người dùng
  - Dễ hiểu đối với con người
  - Có tiềm năng hữu ích
  - Mới
  - Có giá trị trong việc học mô hình
  - Biểu diễn kiến thức nhất định
  - Phục vụ làm sáng tỏ giả thuyết (it validates a hypothesis)

Classification of data mining systems



Phân loại

- Loại dữ liệu
- Loại kiến thức
- Loại phương pháp kỹ thuật
- Loại ứng dụng

Libraries

- Basic libraries for data science
  - NumPy
  - SciPy
  - Pandas
  - IPython notebook
  - matplotlib

Libraries

- Libraries for machine learning
  - scikit-learn
  - Theano
  - Keras
  - Tensorflow

Libraries

- Libraries for NLP
  - NLTK
  - gensim
  - Scrapy
  - Pattern
  - Beautifulsoup

Libraries

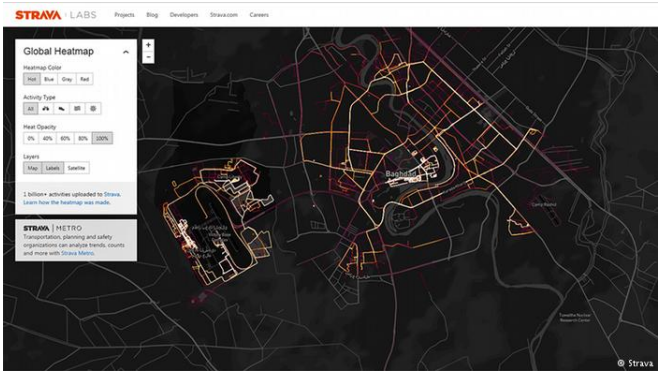
- Libraries for visualization
  - matplotlib
  - seaborn
  - Bokeh
  - Plotly

Libraries

- Libraries for image processing / computer vision
  - scikit-image
  - OpenCV python
  - SimpleCV
  - Tensorflow

Example applications

- Service providers
  - Dữ liệu hành vi người dùng, nguồn dữ liệu di động, nguồn dữ liệu từ app



Example applications

- Retail
  - Chia người dùng thành Recency, Frequency, Monetary (RFM)
  - Phân loại hành vi mua sắm
  - Chiến lược marketing cho các nhóm khách hàng khác nhau

Example applications

- E-commerce
  - Hệ thống bán hàng trực tuyến
  - thegioididong, fpt shop, tiki

Example applications

- Supermarket
  - Sắp xếp gian hàng, chương trình khách hàng thân thiết, tần suất loại hàng theo móc thời gian

Example applications

- Anomaly detection
  - Phát triển dựa trên outliers detection
  - Crime detection, Faulty detection

Data mining on mobile devices

- 90% thời gian sử dụng mobile là sử dụng app
- Góp phần sinh ra dữ liệu theo luật số mũ
- Data: time series, itemsets/transactions, text(free-form), anonymized data, location/geo, mobile data, social network data, email, web content, web clickstreams, images/video, XML data, and music/audio, and Genomics

Data mining on mobile devices

- Mobile devices lưu trữ thông tin cá nhân, sở thích, thông tin địa lý
- Dữ liệu cần cho chiến dịch quảng cáo, đề xuất dịch vụ, xây dựng hệ chuyên gia
- Dữ liệu vô tận
- Nèn tảng triết học cho thiết kế giao diện
- NoSQL là nền tảng lưu trữ lý tưởng cho loại dữ liệu này


Data mining techniques
Classfication
Clustering
Regression
Outer
Sequential Patterns
Prediction
Association Rules


Business Under standing
Data Understanding
Data Preparation
Modeling
Evalution
Deployment

Benefits of data mining

- Thông tin dựa trên tri thức (knowledge-based information)
- Điều chỉnh vận hành và sản xuất
- Xây dựng chiến lược cạnh tranh hiệu quả, giảm chi phí
- Hỗ trợ quá trình quyết định
- Phát hiện thông tin ẩn (hidden patterns), xu thế tiềm năng (potential trends)
- Hỗ trợ người dùng phân tích dữ liệu lớn

Disadvantages of data mining

- Trao đổi / bán thông tin người dùng cho các công ty khác
- Không có giải thuật tốt nhất cho mọi trường hợp
- Nếu dựa hoàn toàn vào kỹ thuật, dẫn đến sai sót về sau

Top data mining algorithms

- Decision trees
- k-means
- Support vector machines
- PageRank
- AdaBoost
- k-nearest neighbor
- Naive Bayes