Data Warehouse & Data Mining

Chapter 1 Introduction

• Why Data Mining?

Data Mining = การขุดเหมืองข้อมูล / ขุดเหมืองข้อมูล

ในปัจจุบัน Data เยอะขึ้นเรื่อยๆ และสามารถเก็บได้

เช่น เซนเซอร์อุณหภูมิ, เซนเซอร์ปริมาณน้ำฝน ⇒ ซึ่งเป็นเครื่องมือที่มีการเก็บข้อมูลไปเรื่อยๆ
ดังนั้น ข้อมูลก็จะมีการเพิ่มจำนวนขึ้นเรื่อยๆ ทำให้ ล้น!!
รวมไปถึง website เช่น สมมติว่า ถ้ามีคนเข้าไปดู website เจ้าของเว็ปไซต์ สามารถที่จะดู Website นั้นได้
สามารถ ทำนายได้ ว่ามีคนไปดู เยอะในส่วนนั้นของเว็ปไซต์

และ อย่างเช่น กล้องฟิล์ม ก็จะไม่ได้เป็นเหมือนแต่ก่อน แล้วที่เป็น กล้องฟิล์ม เปลี่ยนมาเป็น กล้องดิจิทอล

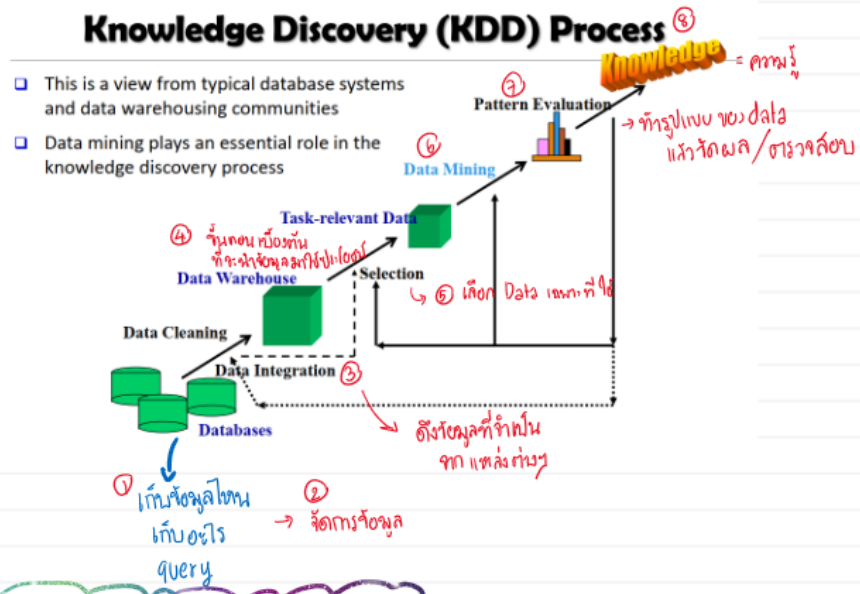
What is Data Mining?

Data Mining = ข้อมูลที่มีมากมาย มาเก็บไว้ แต่มีการ
ผ่านการ สกัด/วิเคราะห์ มาแล้ว

ชื่ออื่นๆ

- → knowledge discovery in databases "การสกัดองค์ความรู้จาก Database"
- → knowledge extraction "การสกัดองค์ความรู้"
- → data / pattern analysis "การวิเคราะห์รูปแบบที่มีในข้อมูล"
- → data archeology "คล้ายๆกับนักขุดโบราณคดี"
- → Business Intelligence "สกัดองค์ความรู้จาก data เพื่อไปใช้กับธุรกิจ"

ภาพรวม Knowledge Discovery (KDD) Process

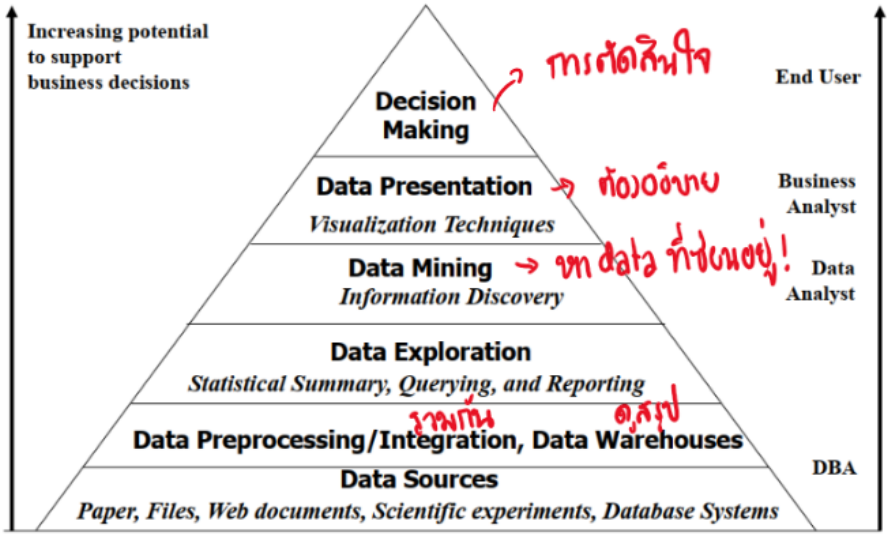


กระบวนการจัดการ Data

Ex การทำ Web mining ?

- Data cleaning
- Data Integration from multiple sources
- Warehousing the data → เป็นข้อมูลที่มีการเก็บไว้ แต่ยังไม่ได้มีการสกัด!
- Data cube construction
- Data selection for data mining
- Data Mining
- Presentation of the mining results —* ต้องนำองค์ความรู้ มาอธิบายได้
- Patterns & knowledge to be used or stored into knowledge-base

Data Mining in Business Intelligence



เนื้อหาหลักๆ 3 เรื่อง

"Data Mining"

1. Pattern discovery = หารูปแบบที่ซ่อนในข้อมูล
2. Classification = จำแนกข้อมูล
3. Clustering = จัดกลุ่มข้อมูล

How the data suppose to look like

Columns (แนวตั้ง) = Attributes, Fields, Features ⇒ คำอธิบายคุณสมบัติของข้อมูล

	id	name	domain_id	closed	city_name	zipcode	geohash	new_open	weighted_average_rating	number_of_chains	...	good_for_groups
0	2	เพรียงหิน ทันตกรรม	2	0	Samut Songkhram	75000	w4rh7g3	0	5.000000	NaN	...	NaN
1	4	Corner House	1	0	Bangkok Metropolitan Region	12150	w4rx73h	0	2.000000	NaN	...	NaN
2	5	วัดโลกยสุธาราม	4	0	Phra Nakhon Si Ayutthaya	13000	w4x96jk	0	4.000000	NaN	...	NaN
3	6	บ้านค้าราโอเกะ	1	0	Bangkok Metropolitan Region	10700.0	w4rqw9q	0	0.000000	NaN	...	NaN
4	7	Buono Caffe	1	0	Bangkok Metropolitan	10220	w4rx4gd	0	3.738462	NaN	...	NaN

→ 1 จุด

row (แถว)

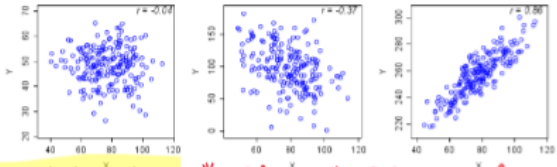
Records, Data point

⇒ ข้อมูลแต่ละตัว

เทคนิค ของ Data mining

1 Data Mining Functions: (2) Pattern Discovery

- Frequent patterns (or frequent itemsets)
  - What items are frequently purchased together in your Walmart?
- Association and Correlation Analysis



- A typical association rule ← * ทำให้คนรู้จัก Data mining!
  - Diaper → Beer [0.5%, 75%] (support, confidence)
  - Are strongly associated items also strongly correlated?
- How to mine such patterns and rules efficiently in large datasets?
- How to use such patterns for classification, clustering, and other applications?

Association rule

ตัวอย่างที่ทำให้คนรู้จัก ⇒ Diaper & Beer
(ผ้าอ้อม)

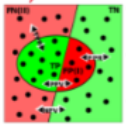
⇒ ใช้เทคนิคนี้ในการวิเคราะห์ ใบเสร็จของคนที่มาซื้อปัง
พบว่า ① คนที่ซื้อผ้าอ้อม มักจะ ซื้อเบียร์ด้วย
② และ ก็พบว่า มีการ สัมพันธ์กันมากกว่า คู่อื่นๆ
③ และ มักพบว่าเป็น พ่อคุณพ่อ

2 Data Mining Functions: (3) Classification

การจำแนกข้อมูล

↳ เลือก 1 Column / 1 attribute เพื่อ เอา Column อื่นๆ มา ทำนาย

- Classification and label prediction
  - Construct models (functions) based on some training examples
  - Describe and distinguish classes or concepts for future prediction
    - Ex. 1. Classify countries based on (climate) ⇒ ข้อมูลเชิงคุณภาพ
    - Ex. 2. Classify cars based on (gas mileage)
  - Predict some unknown class labels
- Typical methods
  - Decision trees, naïve Bayesian classification, support vector machines, neural networks, rule-based classification, pattern-based classification, logistic regression, ...
- Typical applications:
  - Credit card fraud detection, direct marketing, classifying stars, diseases, web-pages, ...

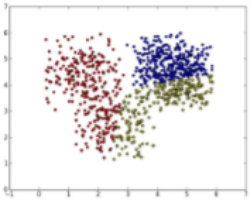


มีค่า y

⇒ เป็นแบบมีผู้สอน

3 Data Mining Functions: (4) Cluster Analysis

- Unsupervised learning (i.e., Class label is unknown)
- Group data to form new categories (i.e., clusters), e.g., cluster houses to find distribution patterns
- Principle: Maximizing intra-class similarity & minimizing interclass similarity
- Many methods and applications



ไม่มีคำตอบให้ ทำนาย
แค่ จัดกลุ่ม
↳ ข้อมูลที่มีลักษณะ: คล้ายกัน