

Data Warehouse & Data Mining

Data Warehouse คือข้อมูลที่เก็บมา นำมาคัดแยกและจัดเก็บ หลังจากนั้นจะนำไปทำ Data mining ต่อเพื่อทำการบันทึกข้อมูลที่เกี่ยวข้องกับตัวข้อมูลหลักในปแต่ละเรื่องๆ คล้ายๆกับการขุดหาเพชร ทำการขุดจนกว่าจะเจอเพชรที่แท้จริง

What Is Data Mining?

- Data mining (knowledge discovery from data)
 - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
 - Data mining: a misnomer?
- Alternative names
 - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything “data mining”?
 - Simple search and query processing
 - (Deductive) expert systems



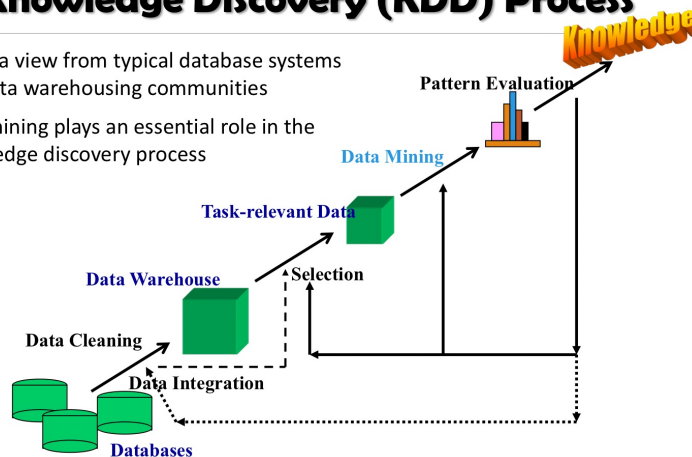
ขั้นตอนการทำเหมืองข้อมูล

ประกอบด้วยขั้นตอนการทำงานย่อยที่จะเปลี่ยนข้อมูลกลายเป็นความรู้ ประกอบด้วย

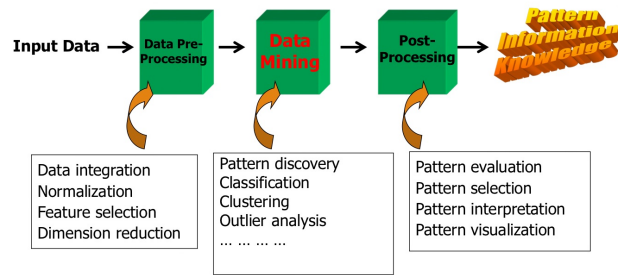
1. Data Cleaning คัดข้อมูลที่ไม่เกี่ยวข้องออกไป
2. Data Integration รวมข้อมูลที่มีหลายแหล่งเป็นชุดเดียวกัน
3. Data Selection ดึงข้อมูลสำหรับการวิเคราะห์จากแหล่งที่บันทึกไว้
4. Data Transformation แปลงข้อมูลให้เหมาะสมสำหรับการใช้งาน
5. Data Mining เป็นขั้นตอนการค้นหารูปแบบที่เป็นประโยชน์
6. Pattern Evaluation เป็นขั้นตอนการประเมินรูปแบบที่ได้จากการทำเหมืองข้อมูล
7. Knowledge Representation การนำเสนอความรู้ที่พบ

Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



KDD Process: A View from ML and Statistics



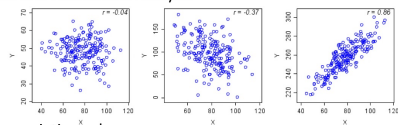
- This is a view from typical machine learning and statistics communities

โดยจะเรียนหลักๆอยู่ 3 เรื่อง คือ

- Pattern discovery (การหารูปแบบที่ซ่อนอยู่ในข้อมูล)

การสกัดหารูปแบบซ่อนอยู่ในข้อมูล โดยใช้เทคนิค Association rule เป็นเทคนิคที่ทำให้คนรู้จัก Data Mining ทำการวิเคราะห์เช่น คนที่จะมาซื้อผ้าอนามัยและเครื่องสำอางมักจะเป็นผู้หญิงวัยรุ่น ซึ่งเจ้าของร้านสามารถนำข้อนี้มาเป็นรูปแบบในการจัดร้านค้า

Data Mining Functions: (2) Pattern Discovery

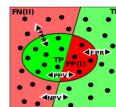
- Frequent patterns (or frequent itemsets)
 - What items are frequently purchased together in your Walmart?
 - Association and Correlation Analysis
- 
- A typical association rule
 - Diaper → Beer [0.5%, 75%] (support, confidence)
 - Are strongly associated items also strongly correlated?
 - How to mine such patterns and rules efficiently in large datasets?
 - How to use such patterns for classification, clustering, and other applications?

-Classification (การจำแนกข้อมูล)

มีการเก็บข้อมูลทุกวัน และมีค่า Y ในการทำนาย

Data Mining Functions: (3) Classification

- 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)
 - 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, ...



-Clustering (การแบ่งกลุ่มข้อมูล)

จัดข้อมูลแบบไม่มีการทำนาย จัดข้อมูลคล้ายๆกันมาอยู่ด้วยกัน

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

