

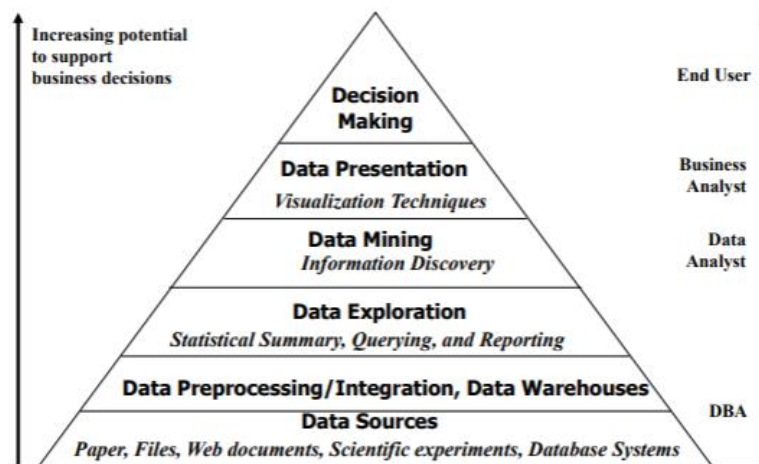
ในส่วนของ Database คือ จัดเก็บข้อมูลอย่างไรให้มีประสิทธิภาพ คิวรีอย่างไร หลังจากนั้นเป็นการจัดการข้อมูลใน multivariate หลังจากนั้นเป็นการทำ Data Integration คือการเอาข้อมูลหลายๆ แหล่งมารวมกันเพื่อไปเก็บไว้ใน Data Warehouse โดย Data Warehouse จะเป็นการดึงข้อมูลที่จำเป็นจะต้องใช้ประโยชน์ มาจากฐานข้อมูลต่างๆ มารวมกัน เพื่อดูข้อมูลที่จะเลือกนำมาวิเคราะห์หาคำตอบที่แท้จริง มาทำ Data mining ซึ่งจะได้เป็น Pattern หรือรูปแบบที่ซ่อนอยู่ในข้อมูล เพื่อทำการวัดผลเพื่อดูว่าข้อมูลที่วิเคราะห์หานั้นเป็นองค์ความรู้จริงๆ หรือไม่ โดยการทำการทดสอบองค์ความรู้ที่วิเคราะห์มาได้ ก่อนที่จะได้มาเป็นองค์ความรู้

ตัวอย่าง : การทำ mining ข้อมูลที่เป็นเว็บไซต์
มีขั้นตอนดังนี้

- Data Cleaning การจัดการข้อมูล
- Data Integration From Multiple Sources การรวบรวมข้อมูลจากหลายๆแหล่ง
- Warehousing The Data เป็นการดูข้อมูล จะยังไม่มีวิเคราะห์ข้อมูลที่ซ่อนอยู่
- Data Cube Construction
- Data selection For Data Mining
- Data Mining เป็นการวิเคราะห์ข้อมูลที่ซ่อนอยู่
- Presentation Of The Mining Results *สำคัญมากในการทำ Data mining เนื่องจากเป็นการนำเอาองค์ความรู้ที่ทำการวิเคราะห์แล้วมาแสดงให้ผู้อื่นได้เห็นและสามารถอธิบายให้ผู้ที่เกี่ยวข้องเข้าใจได้
- Patterns And Knowledge To Be Used Or Stored Into Knowledge-Base

Data Mining In Business Intelligence

Data Mining in Business Intelligence



เริ่มจากแหล่งข้อมูล(Data Sources) ซึ่งจะมาจากแหล่งข้อมูลใดก็ได้ เช่น กระดาษ เว็บไซต์ หรือ Database มาทำเป็นข้อมูลดิจิทัลโดยเก็บไว้ใน Database หลังจากนั้นจึงมาทำ Data Preprocessing/Integration จะเป็นการรวบรวมข้อมูลให้เป็น Data Warehouse เพื่อให้สามารถดูข้อมูลหรือสรุปต่างๆได้ ดึงข้อมูลออกมาดูได้ แต่ไม่เห็นว่ามีอะไรที่ซ่อนอยู่ และ Data Mining จะทำหน้าที่เป็นการหาสิ่งที่ถูกซ่อนไว้ในข้อมูล Data Presentation

เป็นการนำเสนอให้ผู้อื่นเข้าใจองค์ความรู้ที่วิเคราะห์ออกมาได้ Decision Making ขึ้นอยู่กับผู้ใช่ว่าจะนำข้อมูลที่วิเคราะห์มาได้ไปใช้ประโยชน์อย่างไรต่อไป

How The Data Suppose To Look Like

How the data suppose to look like

Columns: Attributes, Fields, Features: ค่าที่ใช้อธิบายคุณสมบัติของข้อมูล

	id	name	domain_id	closed	city_name	zipcode	geohash	new_open	weighted_average_rating	number_of_chains	...	good_for_groups
0	2	เคอริเมท์ ฟันทะกกรม	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	วัดโลกยสุธา ราม	4	0	Phra Nakhon Si Ayutthaya	13000	w4x98jk	0	4.000000	NaN	...	NaN
3	6	เบมเบคาวาโอ เกะ	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

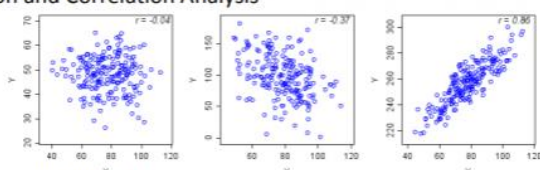
Rows: Records, Data point: ข้อมูลแต่ละตัว

ใน Data ที่จะใช้จะใช้ในการทำ Data mining จะมีลักษณะดังรูปข้างต้น โดย Columns : Attributes, Fields, Features เป็นค่าที่ใช้อธิบายคุณลักษณะของข้อมูล Rows : Records, Data point เป็นข้อมูลของแต่ละตัว

Data Mining Function

Pattern Discovery การวิเคราะห์หารูปแบบที่เกิดขึ้นซ้ำๆในข้อมูล

Data Mining Functions: (2) Pattern Discovery

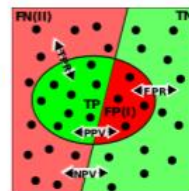
- ❑ Frequent patterns (or frequent itemsets)
 - ❑ What items are frequently purchased together in your Walmart?
- ❑ Association and Correlation Analysis
 - 
 - ❑ A typical association rule
 - ❑ Diaper → Beer [0.5%, 75%] (support, confidence)
 - ❑ Are strongly associated items also strongly correlated?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

โดยเทคนิคที่จะใช้ในการเรียนคือ " Association Rule" โดยเทคนิคที่ทำให้คนส่วนใหญ่รู้จักคือ Diaper & Beer เป็นการวิเคราะห์ที่ไ้เสร็จ โดยสิ่งที่พบคือ คนที่ซื้อผ้าอ้อมมักจะซื้อเบียร์ด้วย ดังนั้นจึงมีการนำข้อมูลที่ทราบไปปรับใช้โดยมี 2 แนวทางคือ 1 จัดให้ผ้าอ้อมและเบียร์อยู่คนละที่ เพื่อในคนเดินผ่านสินค้าชิ้นอื่นด้วย 2 ตั้งเบียร์และผ้าอ้อมอยู่ใกล้กัน เพื่อเป็นการเพิ่มยอดขายให้เบียร์ด้วย ซึ่งทั้ง 2 วิธีต้องมีการเก็บข้อมูลอีกที่ว่า แนวทางไหนให้ผลดีกว่ากัน

Classification การจำแนกกลุ่ม

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, ...



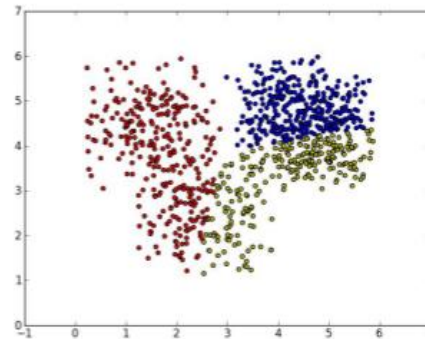
การจำแนกกลุ่มของข้อมูล เช่น การเก็บข้อมูลของกรมอุตุนิยมวิทยา ที่เก็บข้อมูลปริมาณน้ำ อุณหภูมิ ความกดอากาศ ฯลฯ ว่าเป็นอย่างไรของทุกๆ วัน โดย Classification จะช่วยนำเอาข้อมูลอื่นๆ ที่ไม่ใช่ข้อมูลฝนตกของวันนี้ไปทำนายว่าวันอื่นๆ ฝนจะตกหรือไม่ โดยดูจากข้อมูลที่มีอยู่

โดยเป็นการนำข้อมูลต่างๆ มาทำนายข้อมูลตัวหนึ่ง ค่าที่เราอยากทำนาย(Y โดยการนำค่า X มาทำนาย) ไม่เป็นตัวเลข

Cluster Analysis ไม่มีคำตอบให้ทำนาย เป็นการแบ่งกลุ่มเฉยๆ

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



การแบ่งกลุ่ม โดยทำการแบ่งกลุ่มข้อมูลที่มีลักษณะคล้ายๆ กันมาอยู่ด้วยกัน และต่างกันอยู่คนละกลุ่มกัน

**Classification* ใช้ข้อมูล Attributes ที่มีอยู่ ทำนาย Attributes อีกตัวหนึ่งที่เรอยากจะทำนาย

**Cluster Analysis* แบ่งกลุ่มข้อมูล นำข้อมูลที่คล้ายๆ กัน มาอยู่ด้วยกัน