2018年4月13日 星期五

資料探勘 Data Mining

Ch1 Introduction

- **資料探勘**

  1. **WHAT**：在大量的資料中，尋找有價值的資訊或知識。資料為輸入，知識為輸出。

  2. **HOW：步驟**

     - Problem definition 任務理解
     - Data exploration 資料理解
     - Data preparation 資料準備
     - Modeling 知識建模
     - Evaluation 知識評價
     - Deployment 知識部署

  3. **WHY：**

     - Database analysis and decision support

  4. **Funcionalities：**

     - Concept description
     - Association
     - Classification and Prediction
     - Cluster analysis

- **資訊的層級**

  雜訊（Noise）、資料（Data）、資訊（Information）、知識（Knowledge）、智慧（Intelligence）。

- **資料倉儲（Data Warehouse）**

  一個與組織運作資料庫分開維護的決策支援資料庫。將不同來源資料整合起來，找出共通性以協助做決策。

Ch2 Data Warehouse and Online Analytical Processing

- **資料倉儲（Data Warehouse）**

1. 一個與組織運作資料庫分開維護的決策支援資料庫。將不同來源資料整合起來,找出共通性以協助做決策
2. 四個特性
   - Subject-oriented 主題導向
     - 圍繞主題，過濾沒用的資料。
   - Integrated 整合性

     將多個異質資料整合在一起。

     運用到Data cleaning、Data integration。
   - Time-varient 時變性

     儲存不同時間點或期間的資料。

     關鍵結構或多或少都包含時間元素。
   - Nonvolatile 非揮發性

     建置時會將現行操作性系統實體隔開，不互相影響。
3. 三個功用：
   - Information Processing
   - Analytical Processing
   - Data Mining

- **Data Warehouse（DW）與 Operation Database System（ODS）差異**

1. DW主要使用OLAP，用於分析。
2. ODS主要使用OLTP，用於日常作業紀錄。
3.

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

2018年4月13日 星期五

- **Data Cube（資料方塊）**
  1. 介紹：
     - OLAP中最基本的建構單元。
     - 提供快速回應查詢資料的機制。
     - DW子集合所建立的資料集合，由Dimension與Measure所定義的Multi-dimensional Structure，此架構可提供使用者快速而複雜的查詢。
     - ＊A Star-Net Query Model：解決以cube只能表示3個維度之問題，star-net可以表示完整維度。
  2. 可分為三型：
     - distributive(可分散型) Ex：sum、max、min、count
     - algebraic(代數型) Ex：avg、std
     - holistic(整體計算型) Ex：rank、median、mode
  3. 應用：
     - Information processing -support querying
     - Analytical processing - support basic OLAP operations
     - Data mining
- **OLAP（On-line Analytical Processing線上分析處理）**
  1. 介紹：
     - 協助分析者由多個維度觀察變化。
     - DATA CUBE透過OLAP進行操作，OLAP是一種展示multidimensional data的技術，可讓使用者更方便的用不同的面向檢視資料，以分析記錄資訊快速制定策略，常運用在DW前端介面之工具，讓取得、查詢、分析資料時擁有最大的彈性。
  2. 流程：
     - 先建置一個有相關數據分析的Data Cube，提供合適的Dimension與階層。
     - 使用OLAP多維度分析操作,讓不同需求的人對DATA CUBE進行多維度分析操作。
  3. 基本操作模式：
     - Slice(切片)

  - Dice(切塊)
  - Drill-Down(向下擷取)
  - Roll-Up(向上擷取)
  - Pivot(旋轉透視)

- **OLAP與Data Mining 差別**

  1. OLAP：幫助專業經理人驗證假設的工具，讓使用者針對商業問題進行追蹤與探討，透過OLAP工具將分析的結果呈現在使用者面前，以協助做出更合適的決策。
  2. Data Mining：主要分析資料中規律型態，亦即找出資料中的Hidden Patterns，並提出可能性的假設。

Ch3 Data Preprocessing

- **為什麼資料要預先處理（Data Preprocessing）？**

1. incomplete data。資料不完整，可能有缺值。
2. noisy data。雜訊，資料有問題，可能在轉檔時或輸入時錯誤。
3. inconsistent data。不符合，有差異。

- **Data Preprocessing**

1. Data cleaning ：
   - 填補缺值（“unknown”、平均數、最可能的數）
   - 移除雜訊
     1) Binning切割法
        i. equal-depth(frequency):每M份的資料量要相同(數值寬度可不同)
        ii. euqal-width(distance):每M份的數值寬度要相同(資料量可不同)
     2) Clustering
     3) Regression
2. Data integration：資料庫、資料立方體的整合
   - 不同名稱的資料但代表的意思是相同的 Ex.tw=taiwan
   - 不同單位在integration要轉成相同單位
3. Data transformation：標準化，將資料調成固定範圍
   - aggregation/summary Ex.每天轉成每月、每年
   - generalization Ex.年齡轉成年輕、中年、老人
   - normalization：將數值轉換成一個range
     1) min-max normalization :((v-min)/(max-min)) * (new_max-new_min) +new_min
     2) Ex.

byte min=12000,byte max=98000,
xi byte=73600,使用min-max normalization轉換成0-1之間的數值
則新值=((73600-12000)/(98000-12000)) * (1-0)+0=0.71627907

2018年4月13日 星期五

4. Data reduction：減少數量，但達到一樣的分析效果

- 常見方法：

1) dimensionality reduction

   i. 移除不重要的dimensionality

   ii. 有2^d種可能

2) data compression

   i. 通常用於多媒體

3) numerosity reduction

   i. 目地:將數據量變少

   ii. parametric（參數） method：

      a. 包括linear regression、multiple regression、log-linear model

   iii. non-parametric method

      a. histograms

      b. clustering

      c. sampling

4) discretization and concept hierarchy generation:將raw data轉換成higher conceptual level

   Ex.年齡用老、幼、少表示

   i. nominal 無順序 Ex.color

   ii. ordinal 有順序 Ex.rank

   iii. continuous 實數

5. Data discretization（資料離散化）

Ch4 Concept Description：Charaterization and Comparison

- **Data generalization 資料一般化**
  1. 將資料歸納至較高層級。
  2. 方法：
     - Data cube approach 資料方塊法
       1) 具體化一些經常被要求的高成本計算。
       2) Ex.:sum、average、max
       3) 限制：維度的選取還是必須人工手動。
     - Attribute-oriented induction approach 規性導向歸納法
       1) Attribute-removal：沒有更高層級可表示，刪除。
       2) Attribute-generalization：有更高層即可表示，保留並一般化。
          i. 何時停止？超過門檻職（threshold）。
- **Attribute Relevance Analysis 屬性相關性分析**
  1. 為什麼需要？很多屬性有時候是不相關的。
  2. 越能區別類別的，相關性越強。
  3. 方法：
     - Data Collection
     - Analytical Generalization
       1) Decision tree
       2) ID3
          i. Gain(A)=I(p，n)-E(A)
          ii. I(p，n)=(-(a/c)log(a/c))-((b/c)log(b/c))
          iii. E(A)取加權平均
     - Relevance Analysis
     - Attribute-oriented Induction for clads description