Topics in Data Engineering

Session 3

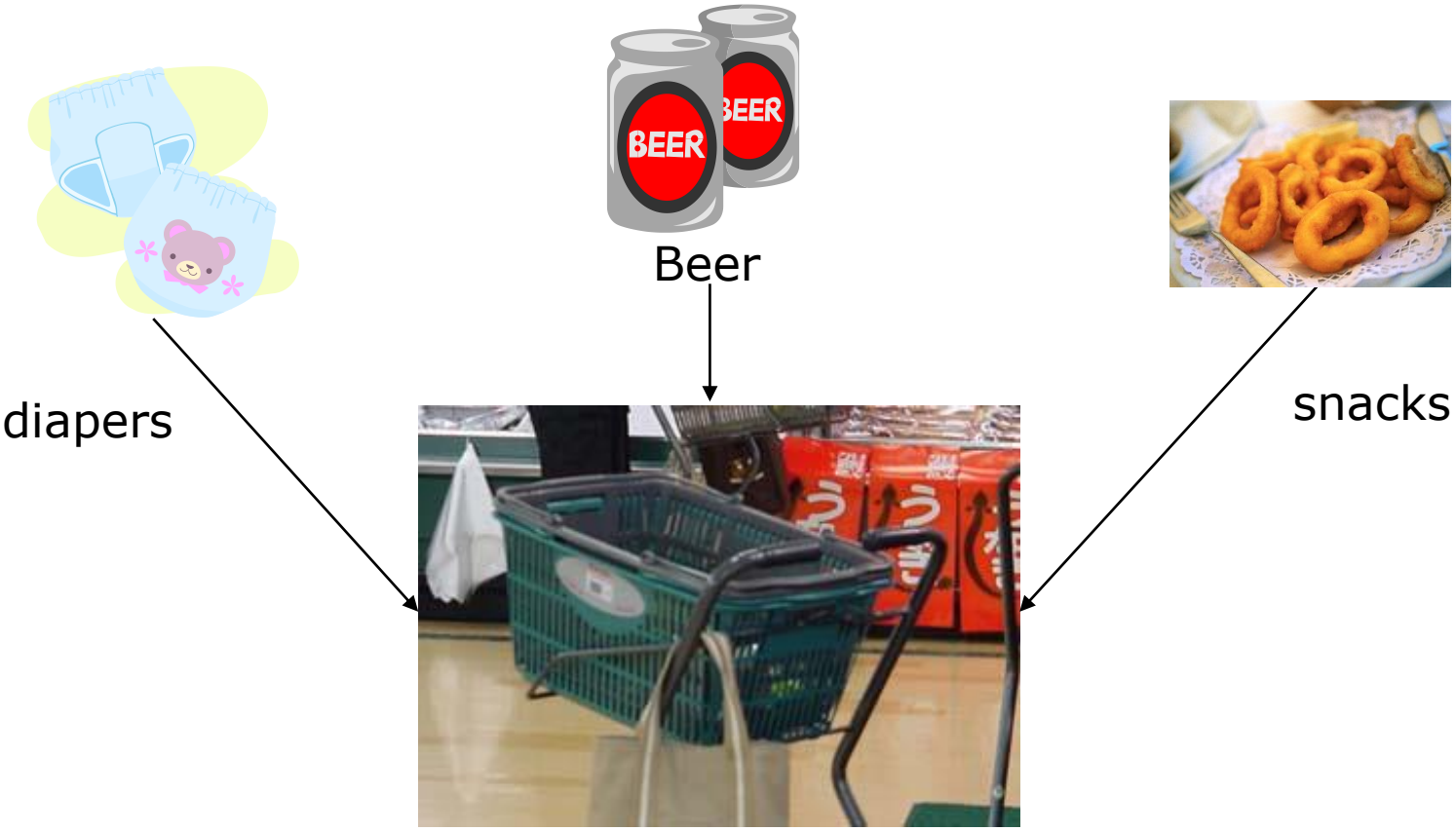
Masaomi Kimura

Topics in this session

- Association analysis
- Memory based reasoning

Association analysis

Association analysis



What we get by this method

- Association rules
  - Event P $\Rightarrow$ Event Q
- Not all of obtained rules are beneficial
  - Bought diapers $\Rightarrow$ bought beer
  - Bought expensive computer
    $\Rightarrow$ contracted 3-year guarantee
  - Bought a USB memory
    $\Rightarrow$ bought sprite

A sample

Customer#	Orange Juice	Coke	Coffee	Cake	Pizza
1	1	1			
2	1		1	1	
3	1				1
4	1	1			1
5		1		1	

Confidence/support/lift

- Confidence= conditional probability of Event A under the occurrence of Event A

$$P(A|B) = \frac{N(A \cap B)}{N(B)}$$

- Support=joint probability of Event A and Event B

$$P(A \cap B) = \frac{N(A \cap B)}{N}$$

- Lift=the ratio of the confidence to the probability of Event A

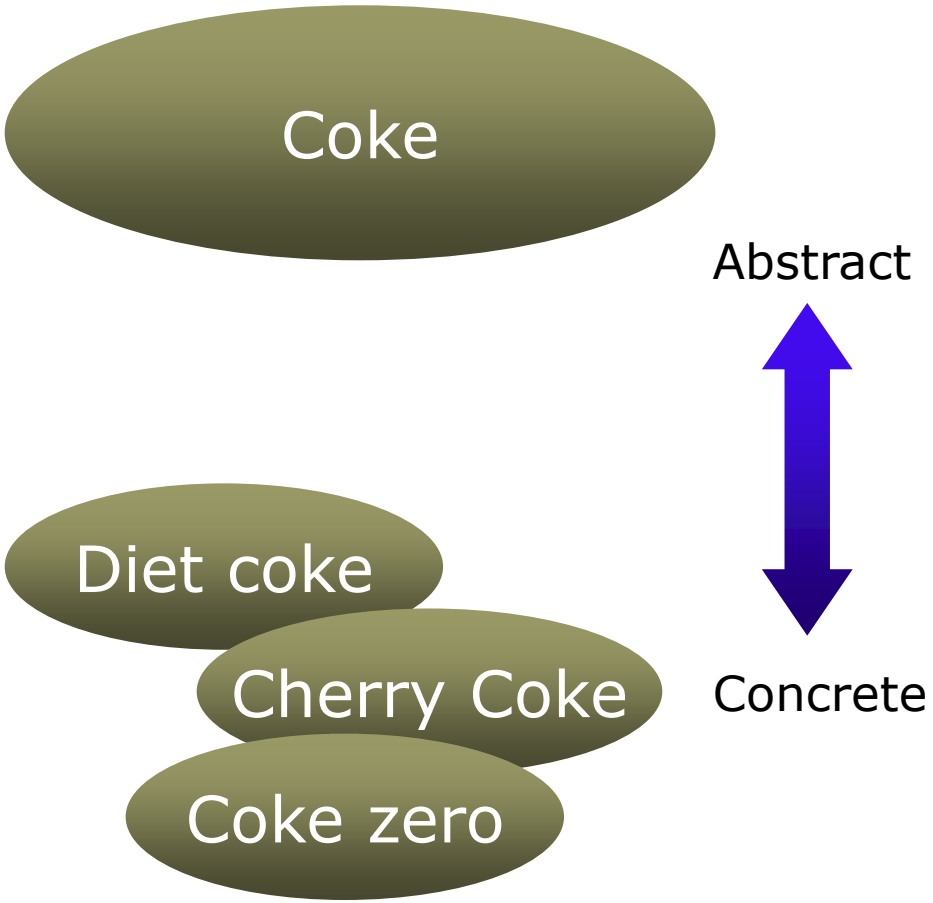
$$\frac{P(A|B)}{P(A)} = \frac{P(A \cap B)}{P(A)P(B)}$$

Procedure

1. Define items/determine abstractness of items
   - Soft drink?
   - Coke, diet coke, coke zero…?

2. Calculate support/confidence/lift
   - Overcome difficulty during the calculation

Choice of correct level for input items

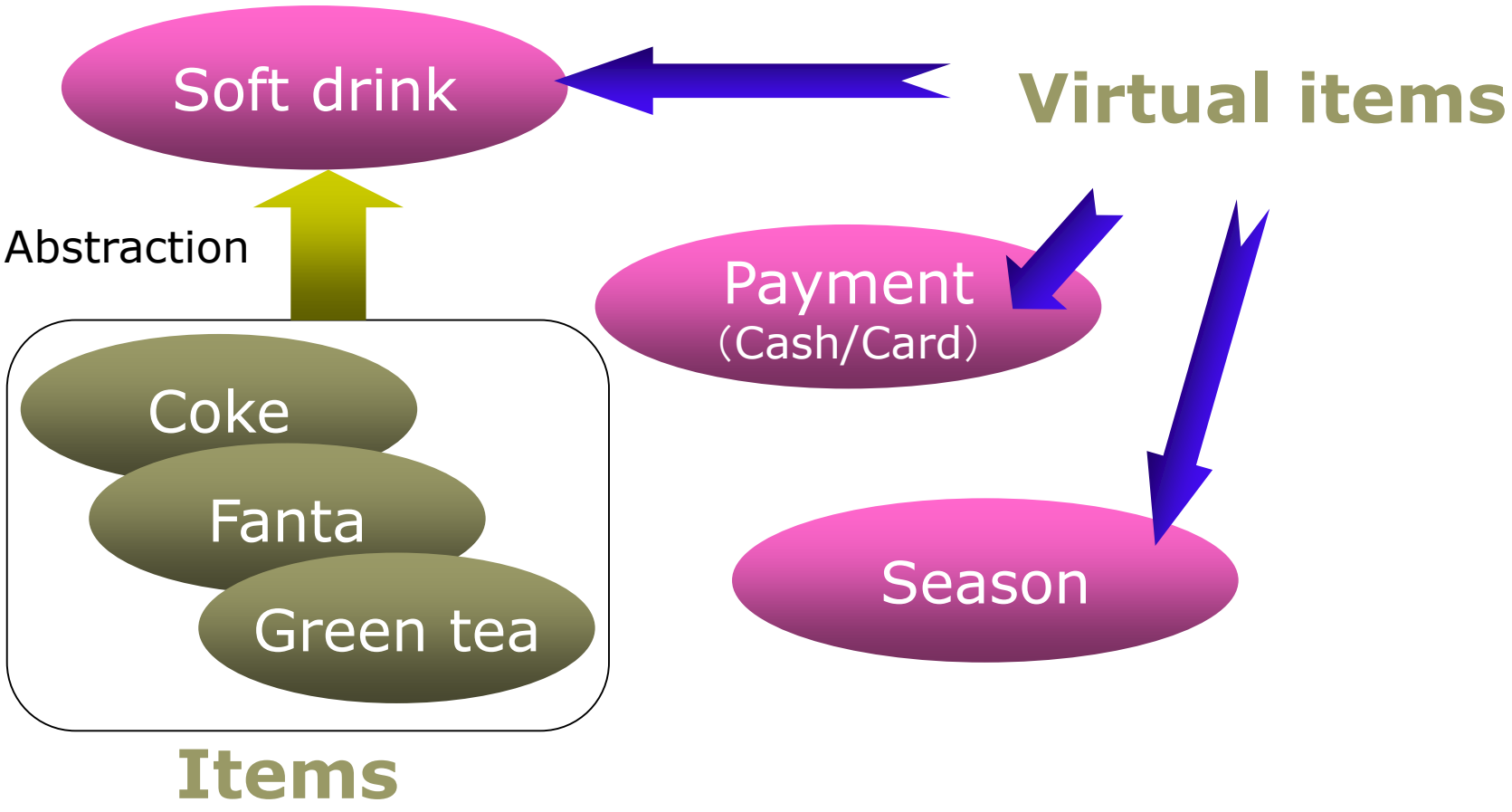


✔ Easy to analyze
✘ ignorance of less freq. event

✔ They make it easy to focus on the particular items

✘ Complex rules are obtained and long execution time is required.

Virtual items



Overcome difficulty during the calculation

- The number of items in an antecedent and an descendant can be more than one.
  - P, Q, …, R $\Rightarrow$ S,T,…,U
  - iPhone $\Rightarrow$ a case, a protect seal
- If there are N items, the rules to be calculated are 2^N!

(Advantage) association analysis

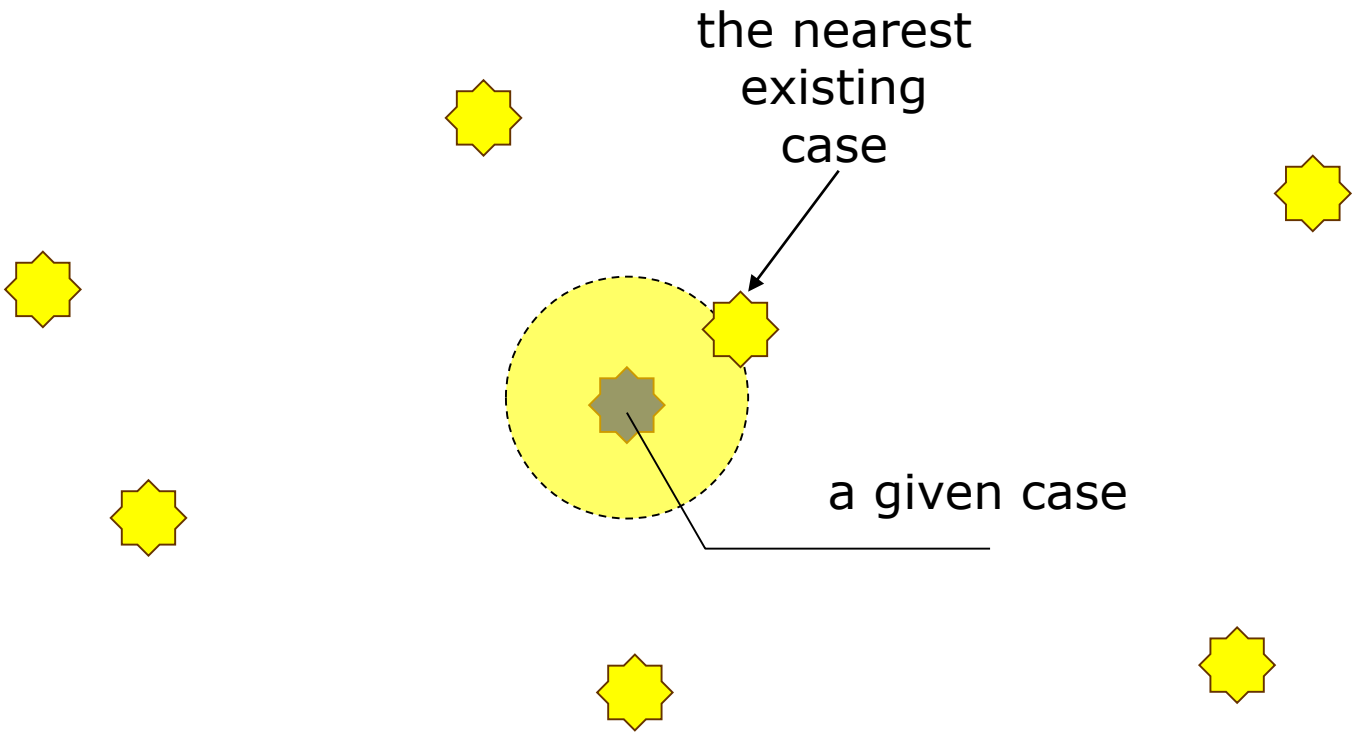
- Generates results easily understandable
- Applicable to variable length data
- uses simple and understandable calculation

(Disadvantage) association analysis

- has a problem of calculation cost, if the number of items is huge
- needs suitable definition/abstraction of items
- does not explain the phenomena of rarely bought items

Memory Based Reasoning

Memory based reasoning



- A distance measure and a combination function are necessary

Procedure

1. Normalize/standardize data in records
2. Search records nearest from an input record
   - a distance function is used to find nearest records
3. Predict a result from the records found in Step 2
   - a combining function is used to find a result from obtained records

1. data standardization/normalization

- Absorb the difference of range size of data
  - normalization
  - standardization

$$z_i = \frac{x_i - \mu_i}{\sigma_i}$$

$$z_i = \frac{x_i - x_i^{\min}}{x_i^{\max} - x_i^{\min}}$$

2. distance function

- A distance function needs to satisfy the following conditions

  Non negativity: $d(\vec{x}, \vec{y}) \geq 0$ (equality stands if and only if $\vec{x} = \vec{y}$)

  symmetry: $d(\vec{x}, \vec{y}) = d(\vec{y}, \vec{x})$

  Triangle inequality: $d(\vec{x}, \vec{y}) + d(\vec{y}, \vec{z}) \geq d(\vec{x}, \vec{z})$

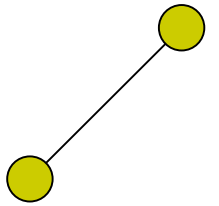
- In practice, some distance function might not satisfy some of the above conditions

Examples of distance functions

- For continuous data, the followings are popular:

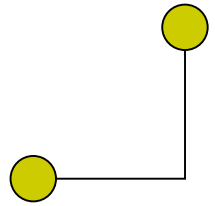
Euclid:

$$d_E(\vec{x},\vec{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$



Manhattan:

$$d_M(\vec{x},\vec{y}) = \sum_i |x_i - y_i|$$



Standardized Euclid

$$d_N(\vec{x},\vec{y}) = \frac{d_E(\vec{x},\vec{y})}{\max_{x,y} d_E(\vec{x},\vec{y})}$$

Examples of distance functions (cont'd)

- As for categorical data, the following distance is popular:

$$e.g. \quad d(male, male) = 0$$

$$d(male, female) = 1$$

$$d(female, male) = 1$$

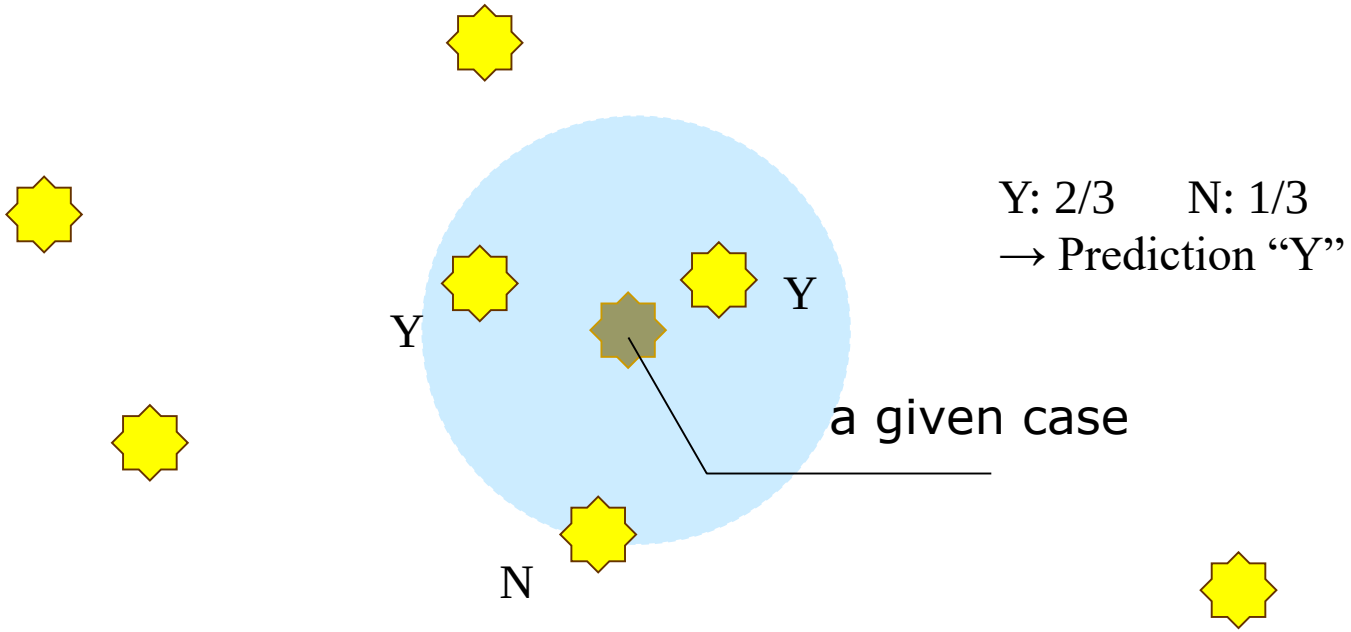
$$d(female, female) = 0$$

Combination of distance functions

- If distance functions both for continuous data and for categorical data are necessary, we need to merge them as a single distance:
  - Summation
  - Euclid
  - Standardized sum

3. a combining function

- Predicts based on the records that have least values of a distance function
- Examples of combining functions



Advantage of MBR

- Advantage
  - Easy to understand why the result was output
  - Applicable to any data type
- Disadvantage
  - Takes long time if the number of existing records is huge