問題発見技法

マーケットバスケット分析

情報学部　堀田敬介

2012/6/19

マーケットバスケット分析とは？

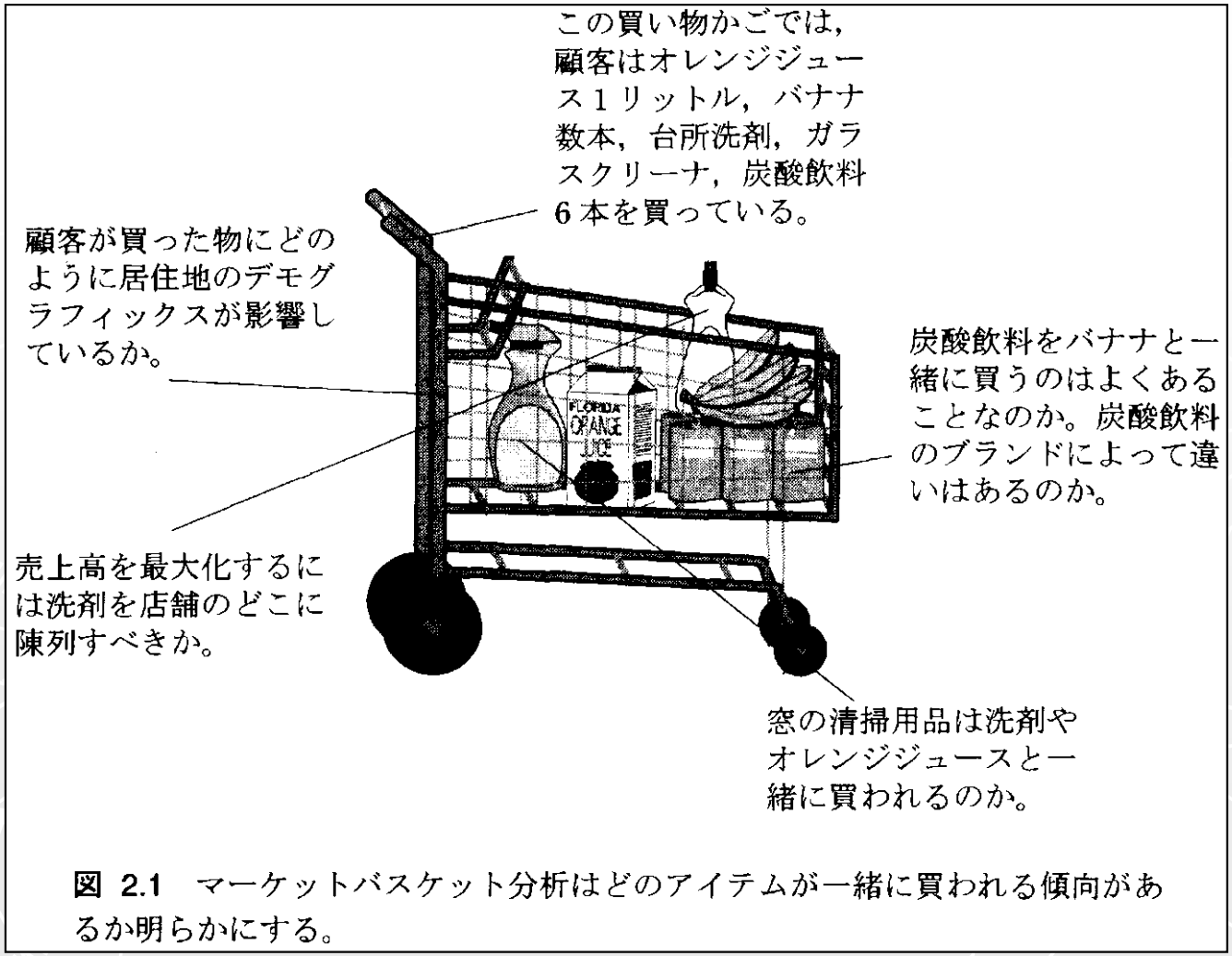


図 2.1 マーケットバスケット分析はどのアイテムが一緒に買われる傾向があるか明らかにする。

〔出展：M.J.A.ベリー「データマイニング手法」p.16 図2.1〕

木曜日に紙おむつを買う若い夫婦は，ビールも買う

マーケットバスケット分析の手順

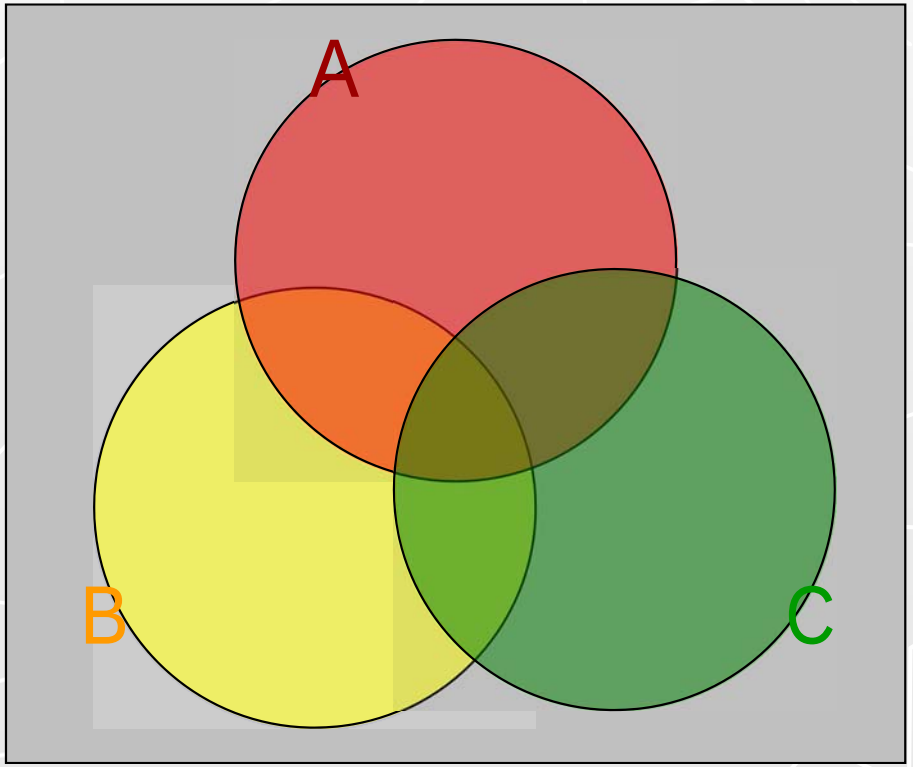
- 基本的な手順
  1. アイテムの適切な集合（水準・内容）を決定
  2. ルールの候補を作る
  3. 調べたいアイテム（orその組合せ）の確率を計算
  4. 信頼性・改善率を計算
  5. ルールの決定

マーケットバスケット分析の手順

- データと出現率
  - 例:3つのアイテムA, B, C

組合せ	出現率
A	45%
B	42.5%
C	40%
A & B	25%
A & C	20%
B & C	15%
A & B & C	5%

全購入者の中で
Aを購入した割合



- Q:どれも買わなかった人は？

マーケットバスケット分析の手順

- ルールの生成

組合せ	出現率
A	45%
B	42.5%
C	40%
A & B	25%
A & C	20%
B & C	15%
A & B & C	5%

A&B の出現率

(A&B)&C の出現率

ルール	if部	then部	信頼性
もし A & B なら C	25%	5%	5% / 25% = 0.20
もし A & C なら B	20%	5%	5% / 20% = 0.25
もし B & C なら A	15%	5%	5% / 15% = 0.33

3回のうち1回はAが(B&C)と一緒に買われる

マーケットバスケット分析の手順

- ルールの生成

ルール	if部	then部	信頼性
もし A & B なら C	25%	5%	5% / 25% = 0.20
もし A & C なら B	20%	5%	5% / 20% = 0.25
もし B & C なら A	15%	5%	5% / 15% = 0.33

信頼性が最も高いルール3を採用したい

But Aの出現率 45% よりも信頼度(33%)が低い！
（即ち，ランダムにAがでるより，ルール3の方が低い）

Aの出現率を加味した尺度「改善率」を求めよう

マーケットバスケット分析の手順

- ルールの生成

ルール	if部	then部	信頼性
もし A & B なら C	25%	5%	5% / 25% = 0.20
もし A & C なら B	20%	5%	5% / 20% = 0.25
もし B & C なら A	15%	5%	5% / 15% = 0.33

組合せ	出現率
A	45%
B	42.5%
C	40%
A & B	25%
A & C	20%
B & C	15%
A & B & C	5%

《改善率の計算例》

$$\frac{P((B\&C)\&A)}{P(B\&C)\cdot P(A)} = \frac{0.05}{0.15\times 0.45} = 0.74$$

$$\left(\text{改善率} = \frac{\text{信頼性}}{P(A)}\right)$$

改善率が1よりも高いルールを採用しよう
（ランダムな判断よりもルールによる予測の方がよいから）

マーケットバスケット分析の手順

- ルールの生成

ルール	信頼性	改善率
もし　A & B　なら　C	0.20	0.50
もし　A & C　なら　B	0.25	0.59
もし　B & C　なら　A	0.33	0.74

改善率が1よりも高いルールはない
（どのルールもよくない）

→ 改善率が1未満の場合，否定にするとよいルールを生む

マーケットバスケット分析の手順

- ルールの生成

ルール	信頼性	改善率
もし　A & B　なら　Not C	0.80	1.33
もし　A & C　なら　Not B	0.75	1.30
もし　B & C　なら　Not A	0.67	1.22

= 1 – (元ルールの信頼性)

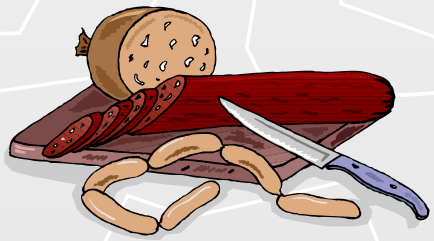
ルール1の否定が最もよいルール

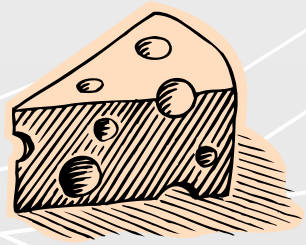
ただし，否定のルールが使えるルールになりえるか？
慎重に判断すべきである.

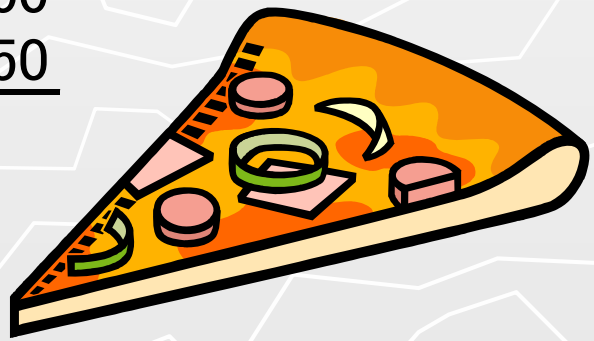
マーケットバスケット分析実施例

- ピザレストランのトッピング

	ピザ	2000
内訳	マッシュルーム	100
	ペパロニ	150
	増量チーズ	200
	マッシュルーム&増量チーズ	400
	マッシュルーム&ペパロニ	300
	ペパロニ&増量チーズ	100
	マッシュルーム&ペパロニ&増量チーズ	200
	トッピングなし	550





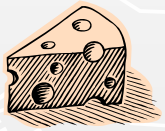


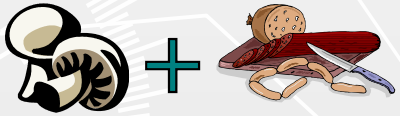
マーケットバスケット分析実施例

- ピザレストランのトッピング
  - 各アイテム，及び組合せの出現率を計算しよう

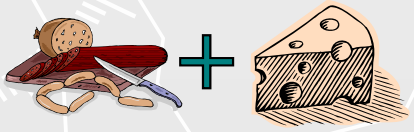
 = (100 + 400 + 300 + 200) / 2000 = 50%

 = (150 + 300 + 100 + 200) / 2000 = 37.5%

 = (200 + 400 + 100 + 200) / 2000 = 45%

 = (300 + 200) / 2000 = 25%

 = 30%

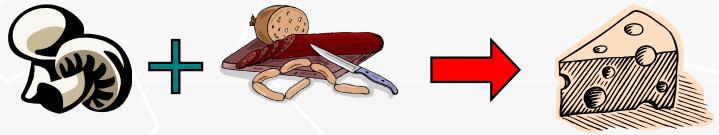
 = 15%

 = 10%

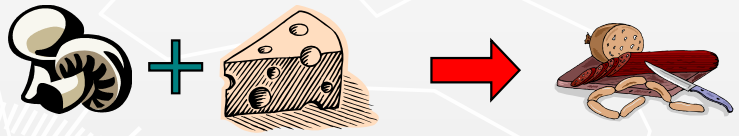
	ピザ	2000
内訳	マッシュルーム	100
	ペパロニ	150
	増量チーズ	200
	マッシュルーム&増量チーズ	400
	マッシュルーム&ペパロニ	300
	ペパロニ&増量チーズ	100
	マッシュルーム&ペパロニ&増量チーズ	200
	トッピングなし	550

マーケットバスケット分析実施例

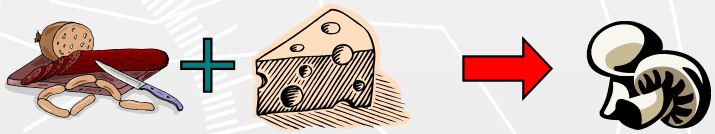
- ピザレストランのトッピング
  - ルールを作り，信頼性と改善率を計算しよう



信頼性: 10% / 25% = 0.4
改善率: 0.4 / 45% = 0.89



信頼性: 10% / 30% = 0.33
改善率: 0.33 / 37.5% = 0.89



信頼性: 10% / 15% = 0.67
改善率: 0.67 / 50% = 1.33

信頼性: 25% / 50% = 0.5
改善率: 0.5 / 37.5% = 1.33

信頼性: 10% / 45% = 0.22
改善率: 0.22 / 25% = 0.89

マーケットバスケット分析の特徴

- 長所
  - 結果が明解
  - 探索的方法である
  - 計算法が単純明解
- 短所
  - 問題の規模に比例して計算量が指数
  - データの属性を限定的にしか使えない
  - 適切なアイテム数の決定が困難
  - 稀にしか購入されないアイテムは説明不能

参考文献

- M.J.A.ベリーほか『データマイニング手法』海文堂(1999)