FUJITSU

shaping tomorrow with you

PWS CUP 2016 の UCIデータおよびコンテストデータ

2016年7月

富士通研究所
山岡 裕司

概要

FUJITSU

- UCI データの分析
- クレンジング
  - トランザクション: Transaction.csv
    - (顧客ID, 伝票ID, 年月日, 時分, 製品ID, 単価, 数量)
  - 品マスタ: Stock.csv
    - (製品ID, 品名)
  - 顧マスタ: Customer.csv
    - (顧客ID, 国)
- コンテストデータ
  - トランザクション: Transaction.csv
  - マスター: Master.csv
    - (顧客ID, 性別, 生年月日, 国)
  - 小規模
    - customer100
    - customer400

FUJITSU

UCI データの分析

UCI データの概要

FUJITSU

- UCI で公開されている Online Retail データセット
  - https://archive.ics.uci.edu/ml/datasets/Online+Retail
- 英国のオンライン店舗での、2010年12月から約1年分の購買履歴
  - 約54万行
  - 7属性 + 品名属性
    - 顧客ID
    - 国（顧客の国）
    - 伝票ID
    - 日時（伝票の年月日時分）
    - 製品ID
    - 単価（当該伝票IDでの製品IDの単価）
    - 数量（当該伝票IDでの製品IDの数量）
  - 主な製品：贈り物
  - 主な顧客：卸売り業者

www.dotcomgiftshop.com
のデータと考えられる。
品名に "Dotcomgiftshop Gift Voucher ..."
などがあるため。

UCI データの一部

FUJITSU

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010/12/1 8:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	2010/12/1 8:26	3.39	17850	United Kingdom
...	...	...	...	...	...	...	...

- 以降、次の構造として考える

顧客ID	国	伝票群
17850	UK	
...	...	...

伝票ID	日時	品群
536365	2010/12/1 8:26	
...	...	...

製品ID	単価	数量
85123A	2.55	6
71053	3.39	6
...	...	...

UCI データの簡易分析結果

FUJITSU

- 簡易分析結果
  - 購買取消の履歴も含まれる
    - 取消伝票の伝票IDには頭に 'C' が付いている
      - 説明には 'c' とあるが、実際は 'C'
  - 3行だけ不良債権調整（Adjust bad debt）があり、伝票IDが 'A' 始まり
  - 数量は負数を含む
    - 取消伝票など
  - 単価は（不良債権調整を除けば）0以上
    - 0は多数あるが、意味不明
    - 0.001は、"PADS TO MATCH ALL CUSHIONS" や "Bank Charges" で使用されているが、実態は0と考える（他の単価は全て小数点以下2位まで）
      - 英国通貨の最小単位は0.01ポンド（ペニー）
  - 顧客IDは空の行がある
    - いわゆるゲストアカウントでの購買か？
  - 国は "Unspecified" を含む
  - 同一顧客IDで国が違う行、同一製品IDで品名が違う行がある

FUJITSU

クレンジング

クレンジング

FUJITSU

- 「購買履歴」として扱い易くするため、次の行を削除した
  - 取消伝票、不良債権調整（それら以外の行の伝票IDは自然数）
  - 単価が0.01未満
  - 顧客IDが空
  - 国が "Unspecified"
  - 製品IDが "BANK CHARGES"
- その後、マスタを切り出し、列順序を変え、日時を分割した
  - 品マスタ (製品ID, 品名)
  - 顧マスタ (顧客ID, 国)
  - トランザクション (顧客ID, 伝票ID, 年月日, 時分, 製品ID, 単価, 数量)

品名、国 は、複数ある場合、
最後に出現した値を使用した

- ヘッダー行を削除し、次のファイル名とした
  - トランザクション： Transaction.csv
  - 品マスタ： Stock.csv
  - 顧マスタ： Customer.csv

Transaction.csv, Customer.csv の概要

FUJITSU

- Transaction.csv の概要
  - 列数: 7
    - CustomerID, InvoiceNo, InvoiceDateYMD, InvoiceDateHM, StockCode, UnitPrice, Quantity
  - 行数: 397,625
    - 削除行数: 144,284
  - 顧客数: 4,333
  - 伝票数: 18,513
  - 年月日, 時分: 2010/12/1 8:26 ～ 2011/12/9 12:50
  - 製品数: 3,663
  - 単価: 0.04 ～ 8142.75
  - 数量: 1～80,995
    - 正数のみになった
- Customer.csv の概要
  - 国数: 36

参考：Transaction.csv の製品ID

FUJITSU

- 特異性が高そうな製品IDとして、次を見つけた

製品ID	品名（Description）
C2	CARRIAGE
DOT	DOTCOM POSTAGE
M	Manual
POST	POSTAGE

FUJITSU

コンテストデータ

フルデータ

FUJITSU

- トランザクション: Transaction.csv
  - クレンジング済UCIトランザクションデータそのまま
    - 詳細は前述
  - 397,625行 7列
- マスター (顧客ID, 性別, 生年月日, 国): Master.csv
  - 顧客IDと国は Customer.csv を使用
  - 性別、生年月日を追加、値は乱数を使い設定
    - 性別: f, m を 5:1 の割合で設定
    - 生年月日: 1930/1/1 ～ 1989/12/31 から設定
      - 年は平均1960 標準偏差10の正規乱数、月日は一様乱数を使用
  - 4,333行 4列
- いずれもヘッダー行なし

小規模データ400

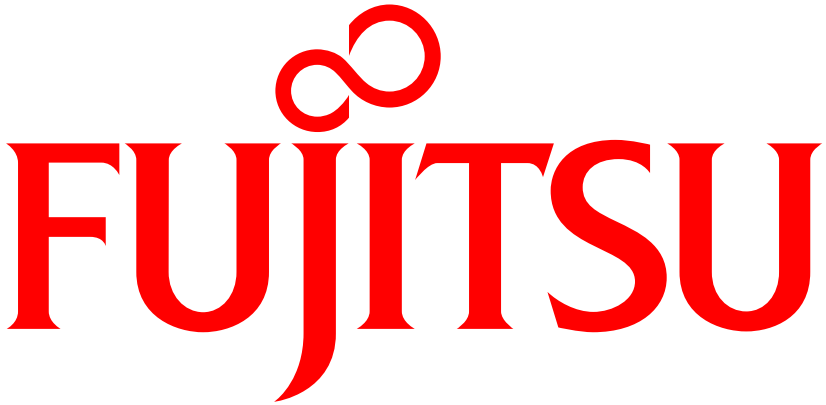
FUJITSU

- 実験や予備戦での使用を想定し、フルデータの一部行を切り出し
- Customer400
  - Master-Customer400.csv
    - Master.csv の最初の400行のみを切り出し
      - ∴ 顧客数: 400
  - Transaction-Customer400.csv
    - Transaction.csv から、Master-Customer400.csv に含まれない顧客IDの行を消去
    - 行数: 38087
    - 伝票数: 1763
    - 年月日, 時分: 2010/12/1 8:45 ～ 2011/12/9 12:50
    - 製品数: 2781
    - 単価: 0.04 ～ 4161.06
    - 数量: 1 ～ 74215

小規模データ100

FUJITSU

- Customer400と同様に、顧客100人データも用意した
- Customer100
  - Master-Customer100.csv
    - Master.csv の最初の100行のみを切り出し
  - Transaction-Customer100.csv
    - 行数: 8695
    - 伝票数: 369
    - 年月日, 時分: 2010/12/1 10:03 ～ 2011/12/9 10:10
    - 製品数: 1878
    - 単価: 0.06 ～ 700.00
    - 数量: 1 ～ 74215



shaping tomorrow with you