

2017年8月9日

ドリトル言語でのデータ処理機能の対応

兼宗 進、小林史弥（大阪電気通信大学）

白井詩沙香（武庫川女子大学）、山澤昭彦（イーテキスト研究所）

Outline

1. 研究背景と目的
2. ドリトルのデータ処理機能
3. まとめと今後の課題

1 新課程でのデータ活用

- 小中高でデータ活用。数学でも統計重視
- 共通教科「情報」でも統計的なデータ処理

情報Ⅰ

- (3) コンピュータとプログラミング
- (4) 情報通信ネットワークと**データの利用**

情報Ⅱ

- (3) 情報と**データサイエンス**
 - (4) 情報システムとプログラミング
-

2 データを扱うプログラミングを提案

- 情報Ⅰ, Ⅱの流れでデータを扱う流れは自然？

情報Ⅰ

- (3) コンピュータとプログラミング
アルゴリズム、シミュレーション
- (4) 情報通信ネットワークとデータの利用
データ操作

情報Ⅱ

- (3) 情報とデータサイエンス
統計処理、グラフ等の視覚化
 - (4) 情報システムとプログラミング
-

3 データを扱うプログラミング

- 情報Iのプログラミングでデータを扱おう
- 複数のデータは配列だが...
- 配列のインデックスを省略できる言語を使う

配列 a の要素を表示する例

<pre>int i; for(i=0;i<5;i++) { print a[i]; }</pre>		<pre>foreach d in a { print d; }</pre>
---	--	--

4 探索アルゴリズムの例

- プログラムといえばアルゴリズム
- 整列は難しいが、探索なら扱いやすい

値が3以下の要素を配列bに格納する例

```
int i, j=0;
for(i=0;i<5;i++) {
    if(a[i]<=3) {
        b[j]=a[i];
        j++;
    }
}
```

```
foreach d in a {
    if(d<=3) {
        b.add(d);
    }
}
```

5 数だけを扱っても嬉しくない！

- 1次元配列だと1種類の値のみ
- 2次元配列だと全体が同じ型！ orz
- 構造体の配列は生徒が理解不能...

配列

30
50
80
70
...

2次元配列

30	50
50	70
80	75
70	80
...	...

データフレーム

名前	数学	英語
田中	30	50
能城	50	70
鎌田	80	75
小松	70	80
...	...	...

6 データベース的なデータ操作

- 言語にデータ操作を入れればSQLが不要に
- 大量のデータから扱いたいデータを取り出す
- データは表形式。2次元配列を簡単に扱いたい

選択

```
tb = ta.selection(price<3);
```

結合

```
tb = ta.join(tc);
```


7 データを扱うプログラミングを提案

- 情報IIのデータサイエンスでプログラミングを扱おう

情報I

- (3) コンピュータとプログラミング
アルゴリズム、シミュレーション
- (4) 情報通信ネットワークとデータの利用
データ操作

情報II

- (3) 情報とデータサイエンス
統計処理、グラフ等の視覚化
 - (4) 情報システムとプログラミング
-

8 データサイエンスとは？

- 膨大な量の構造化データと非構造化データに潜む洞察を、統計、機械学習、データ・マイニング、予測分析などの手法を使用して発見するプロセスのこと^[1]
- ビッグデータ時代の到来により注目を集めている

[1] IBM アナリティクス : <https://www.ibm.com/analytics/jp/ja/technology/data-science/>

9 データサイエンスの能力とは

- データに基づいて課題を解決する能力、統計的な思考力を通じて様々な課題を解決していく能力^[2]

[2] 社会人のためのデータサイエンス入門 : <http://gacco.org/stat-japan/>

10 現行の学習指導要領における取り扱い

- **問題解決のための手段として、「社会と情報」「情報の科学」両教科において簡単なデータ分析が取り扱われている**
- **地球温暖化のような社会問題や学園祭の出し物を決める等、生徒に身近なものが多い**

例) CO₂ 排出量の平均値を算出し、平均値を大きく上回る素材の排気量を明らかにする

11 代表的な分析ツール

- Excel
- R
- SPSS

12 Excel

操作をよく理解している者が、個人で利用する際にはデータの集計や加工、グラフ描画などを行う際に有効だが、操作をよく知らない生徒が授業の中で利用するには問題があり

- 操作の説明が容易ではない
- 処理の順番などを生徒が考えることが難しい

課題 1 操作の説明が容易ではない

- 数式パレットやセルの指定範囲等、初学者の生徒には難しい

(例) Excelで値の平均を求める

- ー リボンから数式パレットを開き、AVERAGE関数を挿入して対象のデータをドラッグして選択
- ー あるいは、セルに「=AVERAGE(A1:A10)」と入力

課題2 処理の順番などを考えることが難しい

- 平均を計算し、次に分散を求めるといった
「データ処理の流れ」という思考になりにくい

⇒ 生徒の思考が Excel の操作手順を考えるといった
「Excel の個々の操作」になってしまう

13 R

- 高度な解析・分析機能を備えており、大学の講義などの教育現場から、実務まで様々な場面で利用
- コマンドラインから操作するCUIの言語であるが、Rstudioを使えばGUIでの利用が可能

⇒ ソフトウェアのインストールが必要で、管理権限の関係で先生方が環境を構築できない可能性あり

14 SPSS

- IBMが提供する有料の統計解析ソフトウェアであり商業用ソフトウェア
- 基本的な解析・分析を扱うことができる
- 高度な解析・分析を行うためには追加でパッケージの購入が必要.

⇒ 高等学校での利用を考えたとき、商業用のソフトウェアは導入の面において敷居が高い.

15 ドリトル

- ドリトルは教育用に設計されたプログラミング言語
- 高等学校の「情報の科学」の教科書や教材で利用されている
- Java の環境があれば特別なインストールは必要なくダウンロードのみで利用可能

16 ドリトル

- 命令を実行した結果に続けて次の命令を送るカスケードと呼ばれる機能が用意されている

かめた=タートル！ 作る。 かめた！ 100 歩く 90 左回り。

ドリトルをデータサイエンス学習に利用することで...

利点1 データの分析の流れをそのままプログラムに書き起こすことができる

利点2 操作手順を伝える場合にはプログラムを表示するだけで良い

17 研究目的

- 高等学校において容易にデータサイエンスを学ぶことができる学習環境を提供することを目的に、ドリトル言語のデータ処理機能を開発する

18 開発した機能

- データサイエンス用の機能を検討する際に、ファイルなどから読み込んだ表形式のデータに対して、次々と処理の命令を送る形でデータの処理や分析を行えるようにした
- カメの登場しないプログラム
- 表をオブジェクトとして扱う

19 開発した機能一覧

処理	命令	動作
データ入力	ファイルから作る	CSV ファイルを読み込む
データ処理	選択 射影 結合 大きい順 小さい順	条件に合致する行を取り出す 列を取り出す 2つのデータをくっつける データを大きい順に並べる データを小さい順に並べる
データ分析	合計 平均 中央値 最大値 最小値	合計を求める 平均を求める 中央値を求める 最大値を求める 最小値を求める

20 開発した機能一覧（つづき）

処理	命令	動作
データ分析	第1四分位数 第3四分位数 度数 標準偏差 分散 クロス集計	第1四分位数を求める 第3四分位数を求める 度数を求める 標準偏差を求める 分散を求める クロス集計を行う
グラフ描画	円グラフ 棒グラフ 折れ線グラフ 散布図 箱ひげグラフ	円グラフを描画する 棒グラフを描画する 折れ線グラフを描画する 散布図を描画する 箱ひげグラフを描画する
表の作成	度数分布表	度数分布表を作る

21 ドリトルのデータ操作 (1)

// 表データの入力

成績=テーブル！ "seiseki.csv" フィルから作る。

// 表データの整列と表示

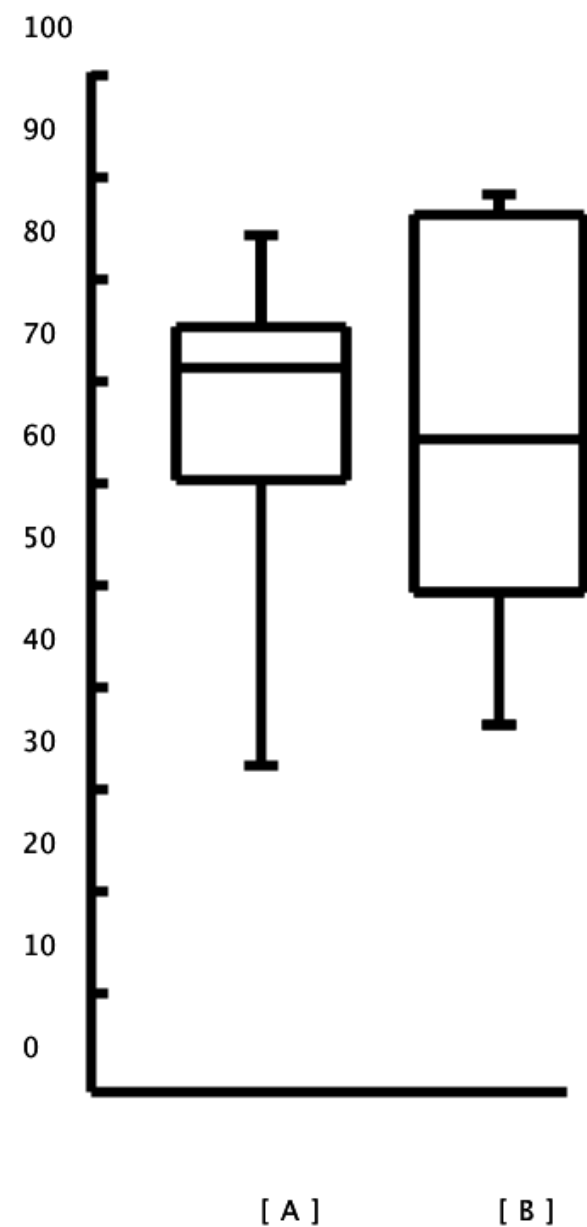
成績！ "組" 小さい順 表示。

// 箱ひげ図の表示

成績！ "組" "国" 箱ひげグラフ 描画。

[番号 組 国 数 理 社 英]

[60 A 74 48 87 71 28]
[59 A 37 56 67 56 85]
[55 A 60 60 67 38 65]
[54 A 74 62 87 31 28]
[51 A 38 50 53 53 75]
[49 A 74 49 92 74 28]
[48 A 75 57 54 66 85]
[44 A 60 53 67 28 65]
[43 A 74 48 65 44 28]
[42 A 75 57 43 54 85]
[38 A 60 55 84 34 65]
[37 A 74 54 38 54 28]
[36 A 75 58 23 54 85]
[32 A 60 62 64 51 65]
[31 A 74 52 31 51 28]
[26 A 39 50 65 64 74]
[25 A 49 48 75 54 57]
[24 A 68 56 37 75 38]
[22 A 59 65 58 61 48]



22 ドリトルでデータ操作 (2)

システム！"db" 使う.

商品=テーブル！"商品.txt"ファイルから作る.

売上=テーブル！"売上.txt"ファイルから作る.

結果1=商品！（売上）結合.

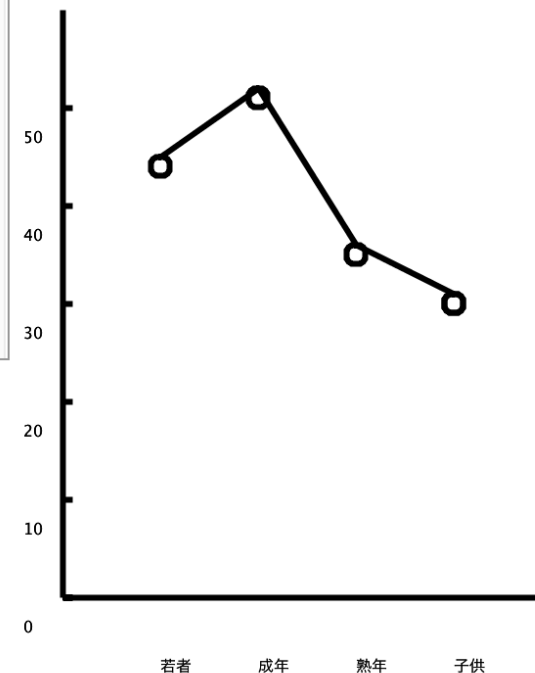
結果1！表示.

結果2=結果1！"年齢" 度数.

結果2！"度数" "年齢" 折れ線グラフ.

結果2！表示.

[G2589 たまごサンド 190g マルタケフード 250 4月4日 水 朝 女 熟年]
[G2879 ハムサンド 190 g マルタケフード 260 4月2日 月 朝 女 熟年]
[G2879 ハムサンド 190 g マルタケフード 260 4月9日 月 朝 女 若者]
[G2879 ハムサンド 190 g マルタケフード 260 4月12日 木 朝 男 成年]
[G2789 ミックスサンド 210 g マルタケフード 270 4月3日 火 夕方 男 熟年]
[G2789 ミックスサンド 210 g マルタケフード 270 4月7日 土 朝 男 熟年]
[G2789 ミックスサンド 210 g マルタケフード 270 4月8日 日 夕方 女 熟年]
[G2789 ミックスサンド 210 g マルタケフード 270 4月10日 火 朝 女 熟年]
[G2762 野菜サンド 200 g マルタケフード 280 4月14日 土 昼 女 成年]
[J4403 コーヒー<微糖> 150ml 丘園製茶 110 4月15日 日 朝 男 若者]
[J4503 カフェオレ 150ml 丘園製茶 110 4月3日 火 朝 女 熟年]
[J4503 カフェオレ 150ml 丘園製茶 110 4月16日 月 朝 男 成年]
[J4503 カフェオレ 150ml 丘園製茶 110 4月16日 月 朝 女 成年]
[J4405 コーヒー<無糖> 150ml 丘園製茶 110 4月7日 土 夕方 男 若者]
[J4405 コーヒー<無糖> 150ml 丘園製茶 110 4月16日 月 朝 男 成年]
[J4408 コーヒー<微糖> 350ml 丘園製茶 135 4月16日 月 朝 女 若者]
[J4406 コーヒー<無糖> 350ml 丘園製茶 135 4月9日 月 夕方 男 成年]
[J4406 コーヒー<武藤敬司> 350ml 丘園製茶 1200 4月9日 月 夕方 男 成年]

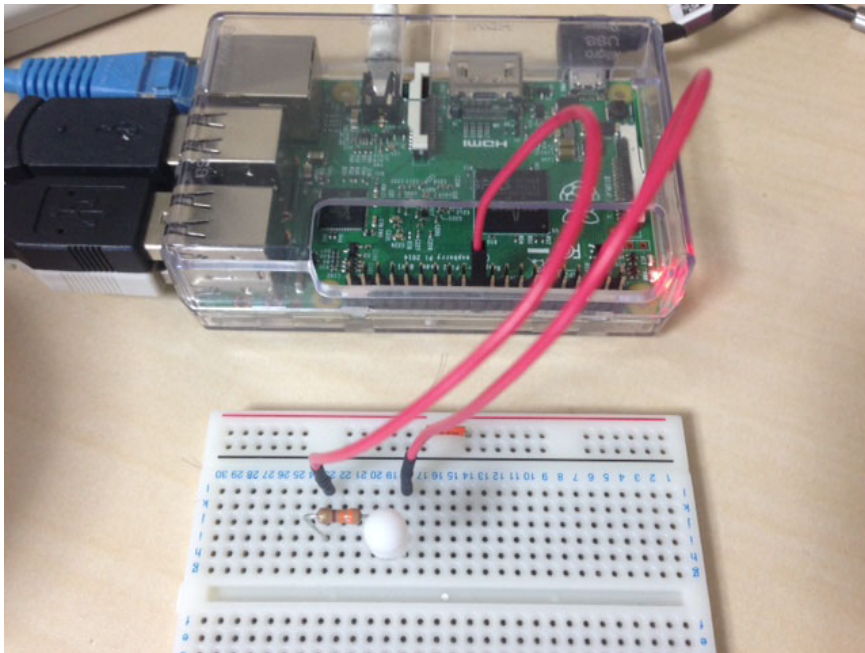


[年齢 度数]	

[若者	45]
[成年	52]
[熟年	36]
[子供	31]

23 計測＋ネットワークの活用も

ドリトル＋ラズベリーパイで計測



24 開発の状況

- 現状版（今日の説明内容）を公開
 - － Webで公開中。V3.10 + zipファイル
 - － 学生が明日ポスター発表
- 今後の計画
 - － 統計関数の充実
 - － 画面でのデータ操作
 - － Web版での動作
 - － Excel形式での出力など

25 作って感じたこと

- インデックスをなくすと配列操作は容易
- 表形式の二次元データが扱えると便利
- DB 的なデータ操作は強力
- プログラミングでの統計的なデータ分析は明瞭
- グラフ表示は心が和む

26 まとめ

- 情報Ⅰ, Ⅱを想定して、「データ処理を中心に置いたプログラミングの流れ」を提案
 - － 情報Ⅰ(3) 探索などデータを扱うアルゴリズム
 - － 情報Ⅰ(4) 表形式の二次元データでデータ操作
 - － 情報Ⅱ(3) 表形式データで統計処理とグラフ表示
 - － 情報Ⅱ(4) サーバのデータをネットワークで操作
- カメやネコのいないプログラミングの世界へ！