第3章　クラスタリングの準備

2008年4月22日(火)

茂木　哲矢

流れ

- ベクトル空間モデル
- 類似度
- 扱うデータセット
- スパース行列

文書クラスタリングとベクトル

文書集合を主題によって分けること

データ解析を行うデータはベクトルで表す

文書をベクトルで表すにはベクトル空間モデルを使う

ベクトル空間モデル

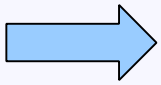
- 文書集合中に現れた単語w_iを第i次元に設定する
- 第i次元の値は文書中のw_iの出現頻度を設定する

$$(f_1, f_2, \ldots, f_n)$$

ベクトル空間モデルの例

d1：私は茨城県にある茨城大学の学生です。
d2：私は茨城県の日立市に住んでいます。
d3：日立市には日立の工場がたくさんあります。

名詞を取り出す



w1：私
w2：茨城
w3：県
w4：大学
w5：学生
w6：日立
w7：市
w8：工場

- 頻度を並べると

(1,2,1,1,1,0,0,0)
(1,1,1,0,0,1,1,0)
(0,0,0,0,0,2,1,1)

単語の重み付け

重みを適切に設定する

- 文書をよく表したベクトルができる
- ベクトルの位置関係とデータの位置関係が近くなる

重みの付け方はTF＊IDF

TF * IDF

- TF（term frequency）

  文書d_i中の単語w_jの頻度（f_{ij}）

- IDF（inverse document frequency）

  全文書数Nをw_jを含む文書数n_jで割った値の対数（$\log(N/n_j)$）

IDF=0だと不都合が生じる

IDF=$\log((N+1)/n_j)$

TF＊IDFの例

- TF＊IDFで重みを求める

0.4055,0.8109,0.4055,1.0986,1.0986,0.0000,0.0000,0.0000
0.4055,0.4055,0.4055,0.0000,0.0000,0.4055,0.4055,0.0000
0.0000,0.0000,0.0000,0.0000,0.0000,0.8109,0.4055,1.0986

正規化

サイズの大きな文書は大きなベクトル

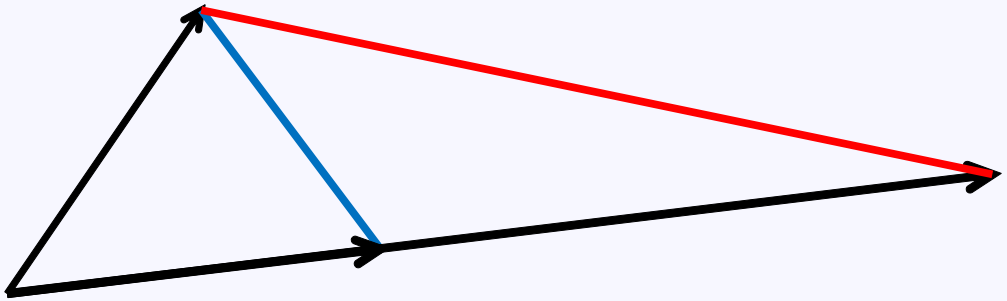
ベクトルの長さを1に正規化する

データ間の類似度

ユークリッド距離

特徴

- 最も単純
- 現実のデータ間の類似度が適切に表現されない



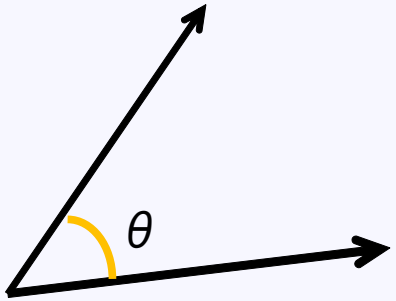
余弦尺度

特徴

- 文書クラスタリングで最も使われる

  内積の定義より

$$\cos\theta = \begin{cases} \dfrac{(a,b)}{\|a\|\|b\|} & (\|a\| \neq 1, \|b\| \neq 1) \\ (a,b) & (\|a\| = 1, \|b\| = 1) \end{cases}$$



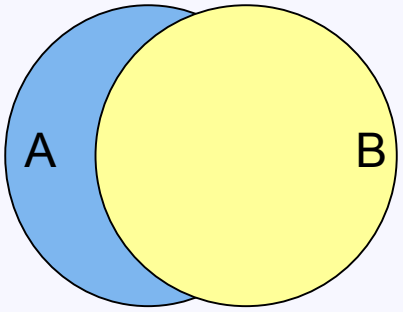
- 正規化されている場合，ユークリッド距離を求めることと余弦尺度を求めることは本質的に同じ

Jaccard係数

特徴

- カテゴリカルな値に使える

$$\frac{|A \cap B|}{|A \cup B|}$$



- n次元ベクトルの場合は

$$\frac{\sum_{i \in I} a_i \cdot b_i}{\sum_{i \in I} a_i^2 + \sum_{i \in I} b_i^2 + \sum_{i \in I} a_i \cdot b_i}$$

扱うデータセット

- iris

  Rに付属しているアヤメについてのデータ

- k1b.mtx・tr23.mtx

  CLUTO付属のデータを変換したもの

スパース行列

特徴

- 大規模
- 要素の大部分が0

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 5 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 \end{pmatrix}$$

スパース行列の記述方法

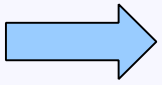
matrix marketの形式

%%MatrixMarket matrix coordinate real general

行列の行数、列数、非ゼロ要素数

非ゼロ要素の行数、列数、その値

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 5 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 \end{pmatrix}$$



```
1 2 1
2 1 2
2 2 1
2 4 5
3 3 1
4 5 8
```

おわり