第3章　グループを見つけ出す（2）

06T4056N　三沢博章

デンドログラムを描く

- デンドログラムの特徴

  - 階層的クラスタリングの結果を参照する際に利用
  - ノードをピラミッドの形で並べて表示する
  - アイテムがどのクラスタに属するかを示す
  - アイテム同士の距離を表示することも出来る

プログラム作成の流れ（1）

①与えられたクラスタの高さを求める

- クラスタが終点ならば高さは1
- 終点以外ならば高さは枝の高さの合計

②ルートノードへの距離の合計を求める

- ノードの深さはそれぞれの枝と深さの最大の数

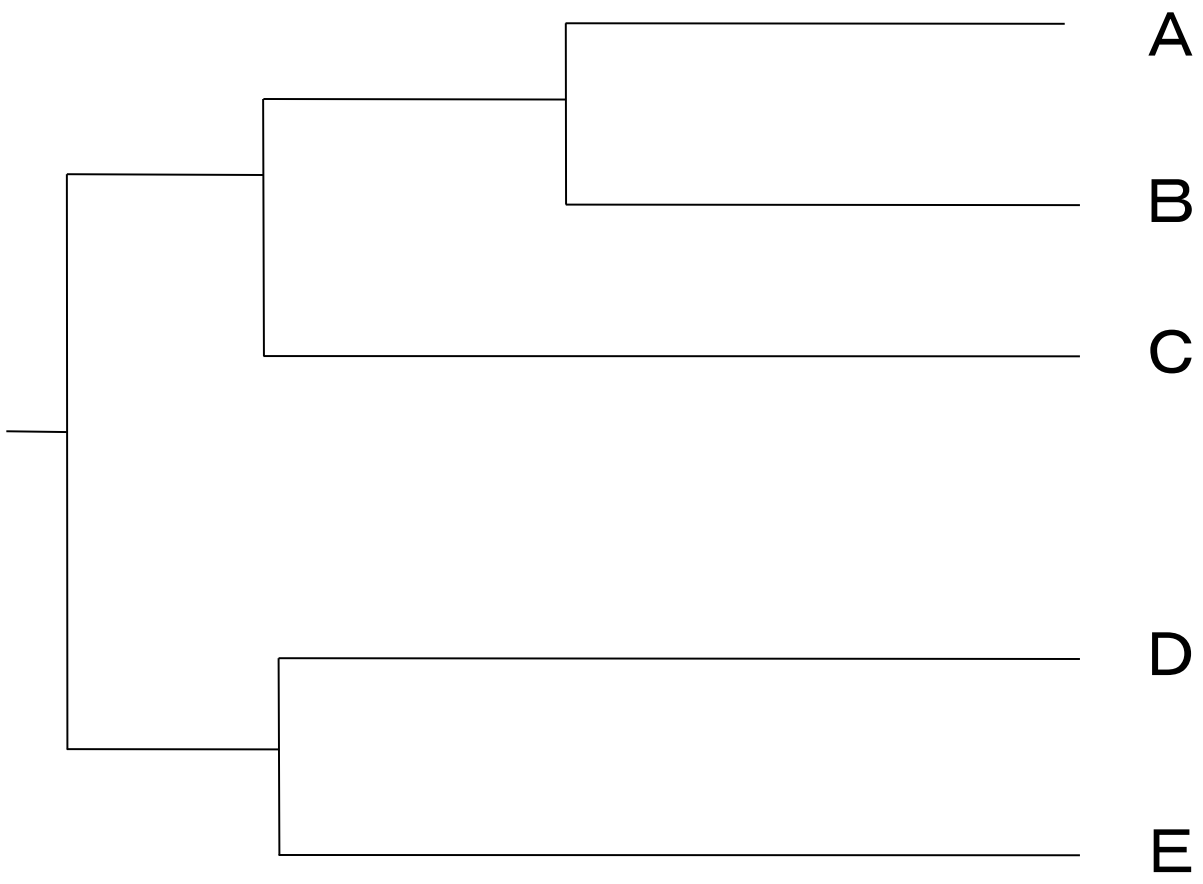
プログラム作成の流れ（２）

③ルートノードを描画する

④終点であればラベル（ブログ名）を描き、
　そうでなければ左右のノードを描く

⑤全て終点になるまで④を繰り返す

図3-1のデンドログラム



列のクラスタリング

行・・・ブログ

列・・・単語

列（単語）が行になるようにデータセットを入れ替える

↓　すると、、、

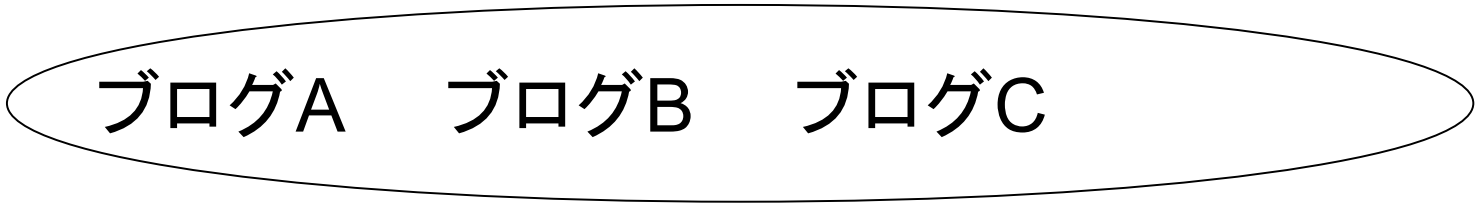
どの単語が一緒によく利用されるか

調べることができる

列のクラスタリング（２）

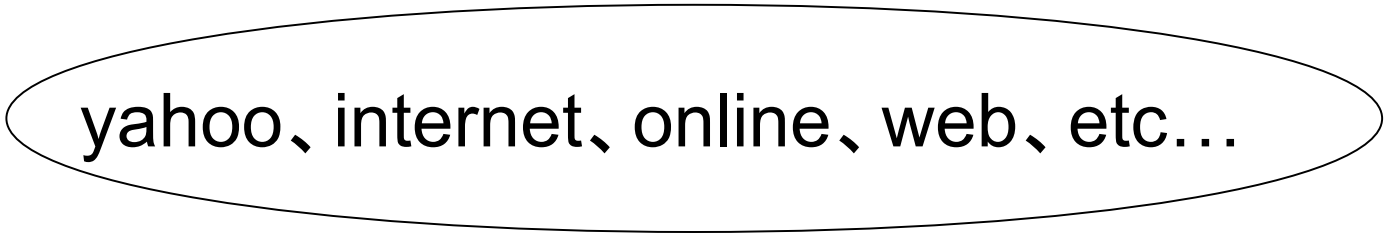
(例)　行＝ブログ、列＝単語の場合

インターネット関連のトピックのクラスタ



行＝単語の場合

インターネット関連のトピックのクラスタ



クラスタリングをする際に

- 変数の数 ≪アイテム数 の場合

意味をなさないクラスタの数が多くなる
可能性がある

（例）
ブログの数 ≪ ブログに使われている単語の数

階層的クラスタリングの欠点

- ツリー形式では、はっきりとしたグループにデータを分けることは出来ない
- 計算量が非常に大きい

↓そこで、、、

K平均法によるクラスタリング

K平均法とは？

- 非階層クラスタリングの代表的な例
- あらかじめ生成するクラスタの数を決めておくことができる

利点：階層的クラスタリングに比べ、かなり高速に動作する

欠点：初期値に依存するため，初期設定により異なる結果となりうる

K平均法のアルゴリズム

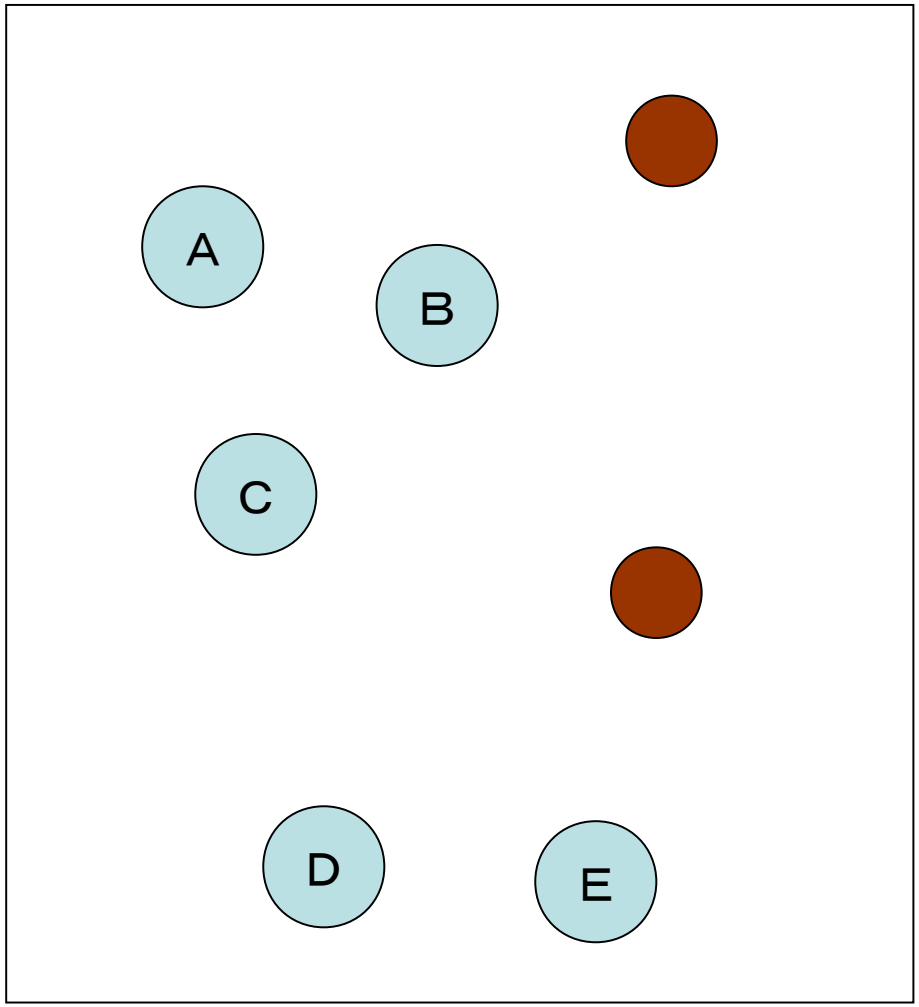
①K個のクラスターの中心をランダムに設定する

②それぞれの個体を最も近い中心に割り当てる

③クラスターごとに中心を計算しなおす

④すべてのクラスター中心が変化しなければ終了
　それ以外は②へ戻る

K平均法で二つのクラスタを作る



K平均法で二つのクラスタを作る