集合知プログラミング
第3章　グループを見つけ出す

04T4027T　斉藤章久

嗜好のクラスタ

・人々から自発的に提供された巨大なデータセットが利用できる

【例】 Zebo
- ・このサイトでは人々がアカウントを作り、彼らが所有しているものや所有したいと考えているもののリストを作成できる。
- ・リストを基に、表現された嗜好が自然にグループ化される様子を見つけ出すことができる。

Beautiful Soup

- Beautiful Soupとは？
  webページをパースし構造化された表現を作り上げる
- ・Beautiful Soupを使うことにより
  type、ID、その他のプロバティでもページのエレメントにアクセスすることができるようになる

  Beautiful SoupのWebページの表現であるスープを作るためには、ページ内容を初期化しさえすればよい
  aのような要素命で呼び出すと、その要素のオブジェクトリストが返ってくる。

Zeboの結果をすくい取る

- ・Zeboのアイテムリストはすべてbgverdanasmallというクラスを持つ
  それにより、ページから重要な情報を抽出することができる
- ・"want"検索のページから最初の50ページをダウンロードし、パースする
- ・10人以上の人が欲しいと思っているアイテムのリストを作る必要がある。そして、無名化されたユーザを列に、アイテムを行にした行列を作成してファイルに書き込む
- ・特定のアイテムをある人が欲しいと思った場合は1、それ以外は0が行列に入っている

距離の基準を定義する

- ・データセットは0か1しかない
- ・tanimoto係数という基準を利用する
  （これは和集合の中での交差集合の率である）
- ・値が1であれば、最初のアイテムを欲しがっている人で2番目のアイテムを欲しがっている人は存在しない
- ・値が0であれば、この二つのアイテムをまったく同じ人々の集合が欲しがっているということである

結果をクラスタリング

- マーケティングに使える情報という観点から見ると驚くような情報は得られない。しかし、はっきりとしたグループは出現している。
- 取得するページ数や初期値を変えることにより様々な結果が得られる

データを2次元で見る（1）

・多次元尺度構成法

アイテムのすべての組の差を用いて、アイテム間の距離がその差の大きさを表すようなチャートを描く。

データを2次元で見る（2）

- すべてのアイテム間それぞれについての目標とする距離を計算する。
- すべてのアイテムは二つのアイテム間の誤差に比例して近づいたり遠ざかったり、少しだけ移動する。
- すべてのノードは自分以外のノードから押されたり、引かれたりする力のの組み合わせにしたがって動く。

クラスタについてその他のこと

これらの他にもできること

・2章のdel.icio.usのデータセットをユーザやブックマークのグループを発見するためにクラスタにする

・ダウンロードできるウェブページの集合はどれでも、単なる単語たちの集合に縮小させること

これらのアイデアは興味深い結果を得るので、様々な範囲に拡張することができる。