重み付きクラスタリング実験

06t4027h 斎藤高行

目的と調査内容

- 入力する単語データを作成し、収集トピック数と重みの値を変え、クラスタリングを実行する

- 正解率と、これら二つの変数の関係性を調べるのが今回の目的である

入力単語データ

- 全ての単語は“分類語彙表”を参考に分類
  - 1.4310-料理
  - 1.3230-音楽
  - 1.4630-機械、装置
  - 1.4210-衣服
  - 1.3374-スポーツ

の5つの分野から各20ずつ(計100)からなる

正解率(1)

- Category_max=10のとき

  w=0.0 : 83%

  w=0.1 : 87%

  w=0.2 : 87%

  w=0.3 : 79%

  w=0.4 : 76%

  w=0.5 : 77%

正解率(2)

- Category_max=30のとき

  w=0.0 : 89%

  w=0.1 : 91%

  w=0.2 : 89%

  w=0.3 : 91%

  w=0.4 : 88%

  w=0.5 : 89%

正解率(3)

- Category_max=50のとき

  w=0.0 : 90%

  w=0.1 : 95%

  w=0.2 : 91%

  w=0.3 : 90%

  w=0.4 : 89%

  w=0.5 : 89%

正解率(4)

- Category_max=100のとき

  w=0.0 : 91%

  w=0.1 : 89%

  w=0.2 : 91%

  w=0.3 : 91%

  w=0.4 : 92%

  w=0.5 : 90%

正解率(5)

- Category_max=150のとき

  w=0.0 : 92%

  w=0.1 : 91%

  w=0.2 : 91%

  w=0.3 : 91%

  w=0.4 : 92%

  w=0.5 : 90%

考察

- 収集トピック数が多ければ多いほど正解率は全体的に高くなる
- 収集トピック数が少ないほど、重みの効果は大きくなる
- 但し、少ない情報からさらに情報を絞るため、大きな重みは逆効果となる場合も考えられる