Recent Developments in Document Clustering

6. Spectral Clustering

6.1 Divide & merge clustering

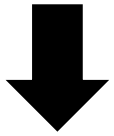
6.2 Fuzzy coclustering

茨城大学工学部情報工学科

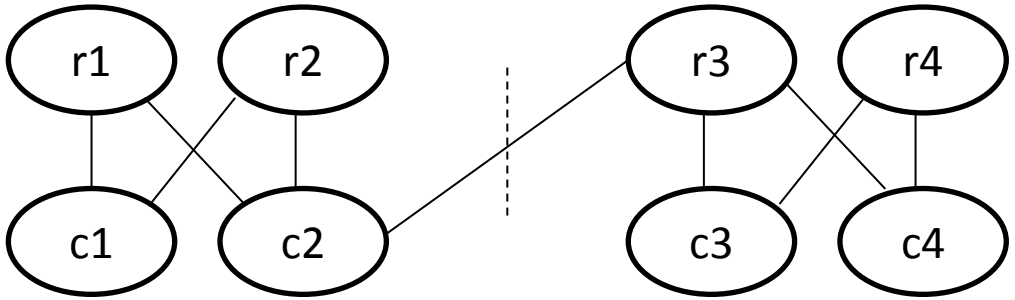
08T4038G 篠塚 晃一

6. Spectral Clustering

- ベクトルモデルはグラフとして解釈できる。



- スペクトラルクラスタリングはこのグラフを分割することで良いクラスタを作る。



6. Spectral Clustering

- 分割の方法
  - ratio cut
  - normalized cut
  - min-max cut

- クラスタがよく分割されるグラフでは全てが同様に動作する。
- クラスタ間にかなりの重複があるときにはmin-max cutが良い分割方法である。

6. Spectral Clustering

- 共クラスタリング(coclustering)の仮定
  - 一緒に現れる単語が同様の概念に関連している。
  - したがって、同様のドキュメントグループを分類するだけでなく、同様の単語グループも分類する。

- 分割を最適化することは行列の特異値分析を計算することと同等である。
- 主成分分析がよく使われる。

6.1 Divide & merge clustering

- 分割を最適化するのはNP-完全問題である。
- アルゴリズムは、２つのフェーズがある。

6.1 Divide & merge clustering

第1フェーズ

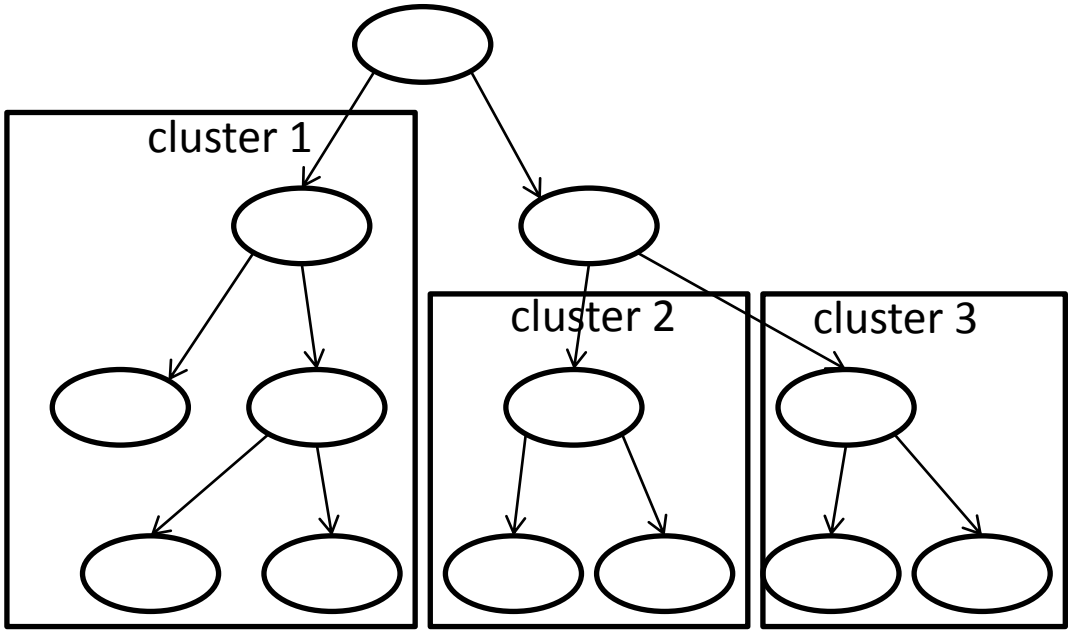
- 階層的クラスタリングは用語とドキュメントの行列から得られるグラフを再帰的に分割することで成り立つ。
- 類似した行列の2番目の固有ベクトルから分割を得ることができるとする。
- べき乗法を使用することで、この2番目の固有ベクトルを概算できる。
- 2番目の固有ベクトルからは分割を最小にするコンダクタンスを見つけることができる。
- この手順は分割することから2つの分割された行列を生じながら繰り返される。

6.1 Divide & merge clustering

第2フェーズ

- 分割フェーズの出力から木のクラスタリングを見つける。
- このフェーズには様々な目的関数を使うことができる。（例：k-平均法）
- 相関関係クラスタリング（correlation clustering)には事前に定義されたクラスタの数によらないという利点がある。
- これはクラスタ数が知られていないインターネット検索エンジンにとって重要な利点である。

6.1 Divide & merge clustering



- 階層構造は繰り返された分割により計算される。
- 下位の木は相関関係クラスタリング（correlation clustering)を使用することで分類される。

6.2 Fuzzy coclustering

- 通常のファジークラスタリングとファジー共クラスタリングの違い
  - 通常のファジークラスタリングでもあいまいな関係を得ることはできるが、ファジー共クラスタリングはデータにそれとなくクラスタに属する度合いを割り当てる。
- 通常の共クラスタリングとファジー共クラスタリングの違い
  - ファジー共クラスタリングでは、あるメンバーシップ関数に従って、前のケースでもクラスタ間の距離があいまいにされること。

6.2 Fuzzy coclustering

- ファジーc-平均法のように共クラスタリングはファジー目的関数を最大限利用する。
- 重要な違いは、「アルゴリズムはドキュメント単独のクラスタよりも用語とドキュメントの共クラスタに最適化される」という集合の概念である。
- 集合は以下のように書くことができる。

$$\sum_{c=1}^{C}\sum_{i=1}^{N}\sum_{j=1}^{K} u_{ci} v_{cj} d_{ij}$$

6.2 Fuzzy coclustering

- そのようなアルゴリズムの例として”Fuzzy Codok”がある。
- ドキュメントに属するものをu, 単語に属するものをv,ドキュメントと単語の相関関係の度合いをdとする。
- m×nの用語とドキュメントの行列を仮定し、ドキュメントのCクラスタと用語のKクラスタがあるとする。
- このとき以下の式でアルゴリズムを更新する。

$$\mu_{ci} = \frac{1}{C} + \frac{1}{2T_u}\left(\sum_{j=1}^{m} v_{cj} d_{ij} - \frac{1}{C}\sum_{j=1}^{m} v_{cj} d_{ij}\right)$$

$$v_{cj} = \frac{1}{K} + \frac{1}{2T_v}\left(\sum_{i=1}^{n} u_{ci} d_{ij} - \frac{1}{K}\sum_{i=1}^{n} u_{ci} d_{ij}\right)$$

6.2 Fuzzy coclustering

- 経験上の評価
  - ファジー共クラスタリングのアルゴリズムはあるデータセットではc-平均法よりも優れている。
- 一般にファジー分割アルゴリズムの弱点
  - 手動であいまいにするパラメータを設定しなければならないこと
  - 最適値はコーパスに基づいてかなり異なるので、データの分配がわからないとパラメータが調整しにくい。

6.2 Fuzzy coclustering

- “Fuzzy Codok”のアルゴリズムと”Soft spectral coclustering”との違い
  - 両方ともドキュメントや用語のためのファジーメンバーシップを生成する。
  - “Fuzzy Codok”は集合の概念を使う。
  - “Soft spectral coclustering”は主成分分析を使う。