# Recent Developments in Document Clustering

Nicholas O. Andrews and Edward A. Fox

## 1. Introduction

新納浩幸

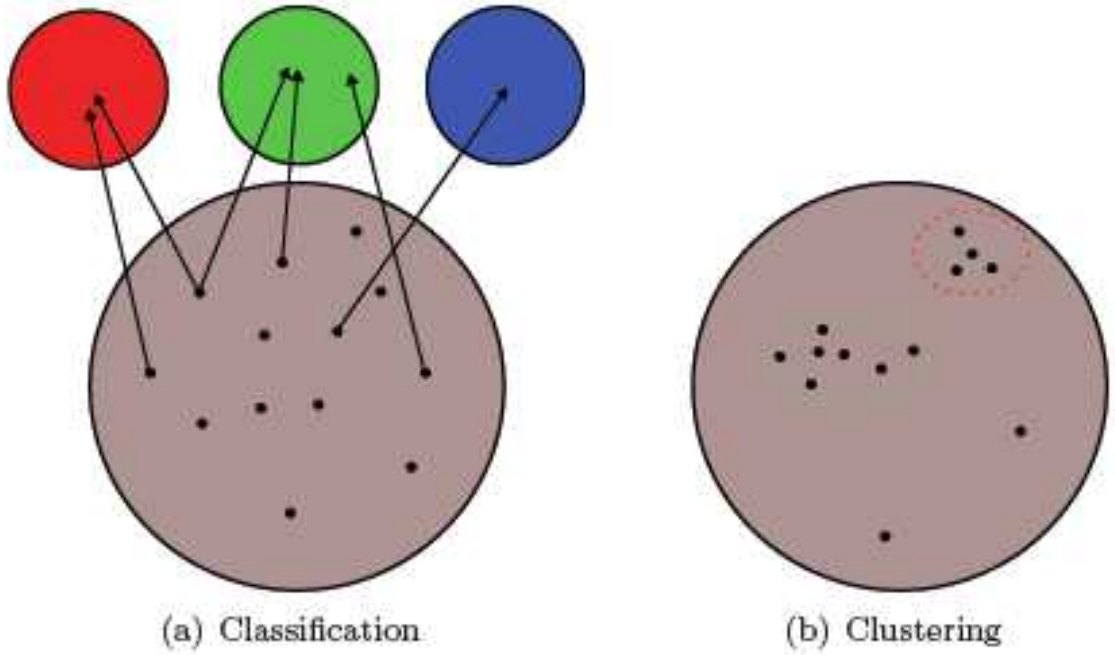
# 文書クラスタリングとは

- データクラスタリングの1つ
- 文書の集合をトピック毎のグループに分割
- 教師なし学習の1つ
- 情報検索、自然言語処理、機械学習などの分野で利用

# Classification と Clustering

Classification は識別、教師有り学習

Clustering はクラスタリング/分類、教師なし学習



(a) Classification

(b) Clustering

# クラスタリングの評価

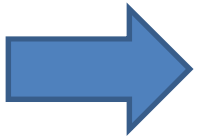
- クラスタ内のデータは類似し、クラスタどうしは類似していないクラスタリング結果が良い
- Overlap を満たすクラスタリングもある（Fuzzy Clustering, Soft Clustering）
- よいクラスタリングのためには、クラスタリングのアルゴリズムと結果の評価法が大事

→ ２章

# Document Model

クラスタリングの最初に考慮すること
文書をどのように表現するか？

ほとんどの場合、ベクトル空間モデルを利用



3章

# Discriminative Type と Generative Type

クラスタリングのアプローチ

- Discriminative Type （4章）

  ある目的関数を設定し、その関数を最適化するようにクラスタリングする

- Generative Type （5章）

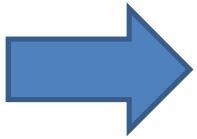
  データの分布を仮定して、その分布に最も合うようにクラスタリングする

# グラフ理論からのアプローチ

索引語文書行列は隣接行列

グラフで表現できる

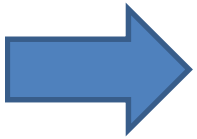
クラスタリングはグラフのカットを求める問題



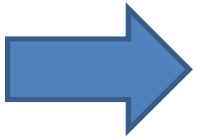
６章

# 次元縮約からのアプローチ

索引語文書行列は巨大な次元



より低次元で表現



7章

# 発展と特徴

- ベクトル空間モデルは単語の出現順が考慮されない
  - → 単語の共起のカウント
- 作成されたクラスタにラベルを付ける
- 本レポートで紹介するアルゴリズムの特徴は9章