jh130050-DA03

佐藤　一誠（東京大学）

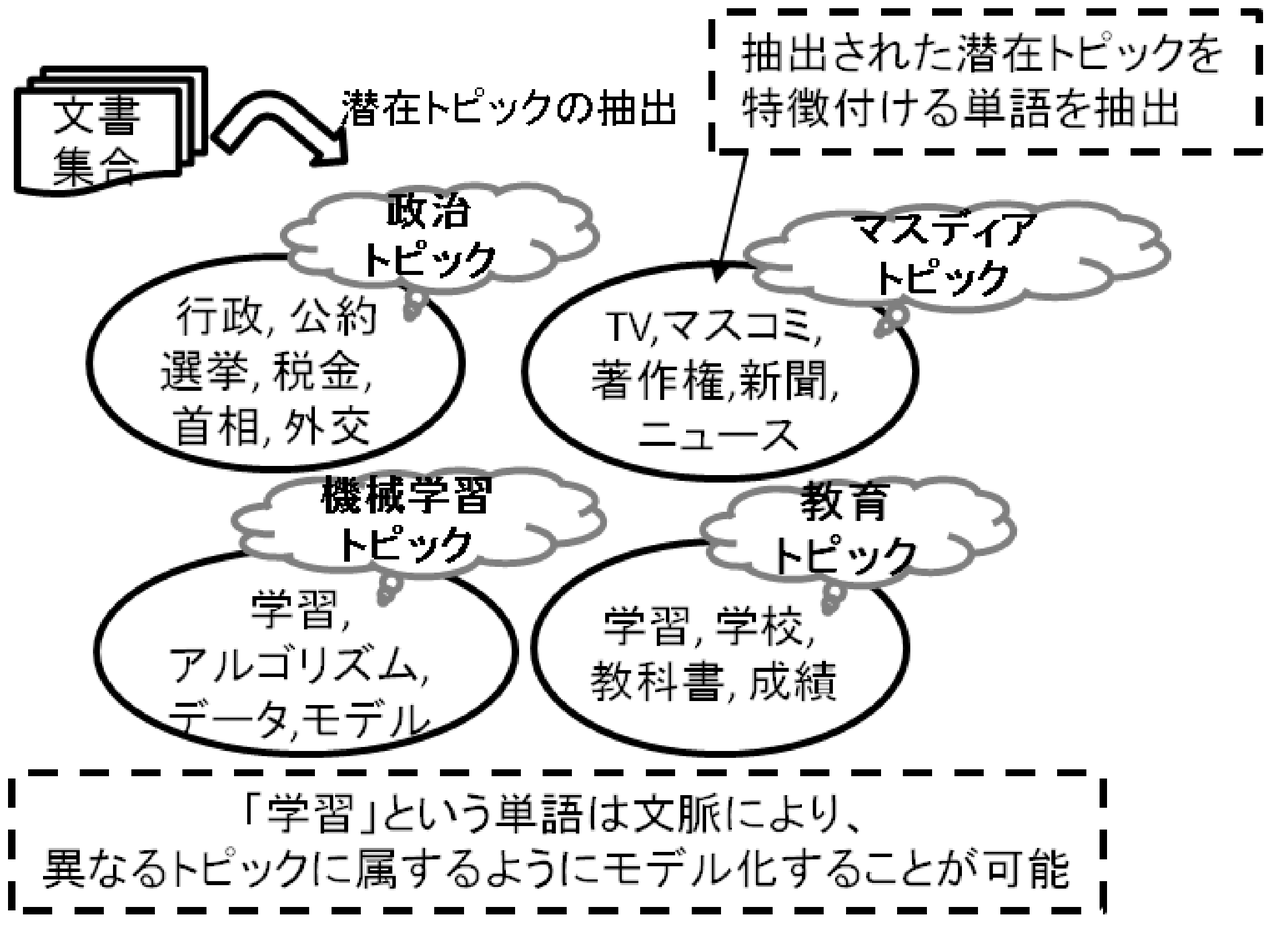
確率的潜在変数モデルの大規模学習アルゴリズム開発

研究背景

潜在変数と呼ばれる確率変数を導入することで、データ中に隠れた情報を抽出する確率的潜在変数モデルの研究がデータ解析において幅広く用いられている。

確率的潜在変数モデルの学習は、非観測の潜在変数をデータから推定する教師なし学習であるため、教師データが不要であり学習データ作成コストが低く、大規模なデータを学習データとして用いることができる。

しかし、確率的潜在変数モデルの学習は非観測の潜在変数を含むため学習アルゴリズムを大規模データで学習するためのアルゴリズム開発が重要なテーマとなっている。



- 教師情報を与えずに、文書をいくつかの潜在トピックに分類する場合を考える
- 潜在トピックモデルは、文書中に内在する隠れたトピック情報を潜在変数として文書の生成過程を学習することで、文書を自動的に分類する

研究概要

対象とする潜在変数モデル:
- Latent Dirichlet Allocation (LDA) [Blei+,03]

対象とする学習アルゴリズム:
- 周辺化Gibbs sampler [Griffiths+04]
- 周辺化変分ベイズ法 [Teh+07,08]

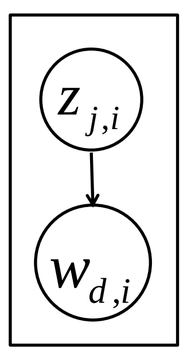
問題点:
メモリ:　O(TN) → O(TV)　本研究
並列化が難しい→容易

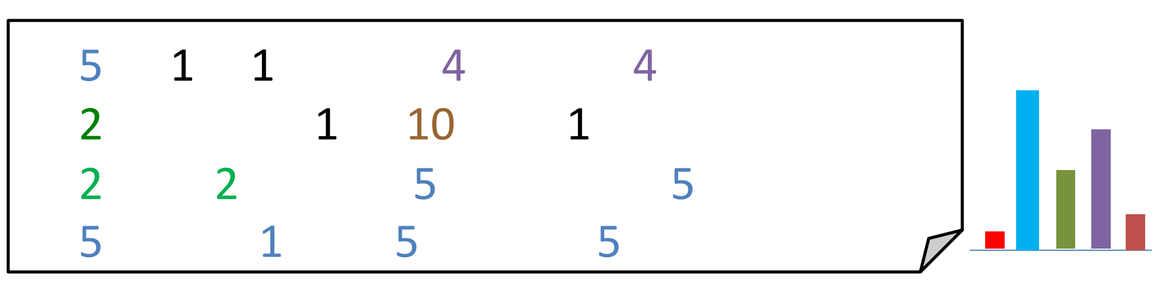
T:潜在変数の次元
N:文書中の全単語数
V:語彙数

潜在変数の仮定:単語は各々トピックを持つ

The apple forms a tree that is small and deciduous, reaching 3 to 12 metres (9.8 to 39 ft) tall, with a broad, often densely twiggy crown.

Apple is an American multinational corporation that designs and sells consumer electronics, computer software, and personal computers.





周辺化変分ベイズ法の学習則

$$q(z_{d,i}=k) \propto \frac{\mathbb{E}[n^{-d,i}_{k,w_{d,i}}] + \mathbb{G}[\beta_0 \tau_{w_{d,i}}]}{\mathbb{E}[n^{-d,i}_{k,.}] + \mathbb{G}[\beta_0]} (\mathbb{E}[n^{-d,i}_{d,k}] + \mathbb{G}[\alpha \pi_k])$$