トピックモデルを用いたグラフ表現に対する

潜在的意味解析に関する研究

Studies of Latent Semantic Analysis on the Graph Representation with Topic Models

学籍番号：201621631
氏名：野沢　健人
Kento NOZAWA

実サービスのデータには，直接観測できない関係性や意味が潜んでいる．このような潜在的意味をデータから抽出することでより深いデータの解析や推薦を行うことが可能になる．本論文では，特に汎用的なデータ構造であるグラフ上での潜在的意味解析を扱う．この問題に対するアルゴリズムとして，トピックモデルの 1 つである Latent Dirichlet Allocation を用いた重複コミュニティ抽出と短文書からのトピック抽出に関する研究を行う．

重複コミュニティ抽出とは，ネットワークから密な部分グラフを抽出する手法であり，例えば，論文をノードとする引用グラフであれば，特定の話題に関する論文の集合を教師なしでまとめることができる．本研究では，大規模グラフを対象とし，サンプル近似による高速なアルゴリズムの挙動を実験的に示すことで，自然言語のデータでしか有効性が明らかにされていないアルゴリズムについて，グラフデータに対する有効性を明らかにする．実験結果から，パラメータを変化させた際のスケーラビリティと抽出精度のトレードオフの関係性を示す．また 6000 万ノード，18 億エッジからなる大規模ネットワークに対しても，既存手法と比較して高速なコミュニティ抽出を実現できることを示す．

短文書に対するトピック抽出は，インスタントメッセージアプリやソーシャルメディアの普及によって顕著になっている問題である．文書からトピック抽出を行うモデルの 1 つである LDA は，1 つの文書の共起情報から潜在変数を推定する．このため，共起情報の限られた短文書に対しては，トピック抽出の性能が損なわれる問題が知られている．本研究では，まずグラフを構築し，その上で行うランダムウォークからなる疑似文書を生成する．この疑似文書を用いることで，より優れた潜在的意味の推定を目指す．実験から既存の擬似文書を用いた手法より少ない共起情報を元にしても高い UCI topic coherence をもつトピックが抽出できることを示す．

研究指導教員：手塚　太郎
副研究指導教員：若林　啓