九州大学学術情報リポジトリ
Kyushu University Institutional Repository

Document Topic Modeling Methods Based on Context Information

李, 文博

https://hdl.handle.net/2324/4496069

出版情報：九州大学, 2021, 博士（情報科学）, 課程博士
バージョン：
権利関係：

（別紙様式 2）

氏　　名　：李　文博

論 文 名　：Document Topic Modeling Methods Based on Context Information
（文脈情報に基づく文書トピックモデリング手法）

区　　分　：甲

論　文　内　容　の　要　旨

ビッグデータ時代とも称される近年，ソーシャルネットワークサービスなどの普及や発展などのために文書データが増大している．入力された文書集合からその背景にあるトピック確率分布を確率的生成モデルで推定するトピックモデリングは，潜在的な意味を理解する手段として，約 20 年間にわたり大きな成功を収めてきた．文脈は，対象となる単語の前後に位置する単語群や，文書系列中において対象文書の前後に位置する文書群として定義され，その活用法はトピックモデリングにおける主要課題である．本論文は，これら 2 種類の文脈の活用に関して，次に列挙する 3 個の貢献を行った．

まず，時間順に並んだ文書系列から，前後に位置する指定数の文書群を文脈として活用して，対象文書のトピック確率分布を推定する問題に対し，トピックの複雑な変遷を推定できる新規手法を提案した．先行研究は，トピック確率分布の文脈への依存性に関する仮定が単純であるため，主要トピック，新興トピック，孤立トピックすべてを正確に推定することが困難である．提案手法は，これら 3 種類のトピックを推定する確率変数を確率的生成モデルに導入し，新規に考案したマルコフ連鎖モンテカルロ法でパラメータ群の値を推定することにより，この問題に対処した．英語ニュース，中国語ニュース，英語医学論文概要，中国語ブログ，英語ツイート，人工データなどの文書データを用いた実験の結果，提案手法は，先端手法を含む比較手法をトピックモデリングと新興・孤立トピック検出の正確性で上回り，ハイパーパラメータへの依存性も低いことが分かった．

次に，入力された文書集合において，前後に位置する指定数の単語群を文脈として活用して，対象となる出現単語が所属する単一トピックを推定するトピック区間分割問題に対して，多様な区間長にも対処できる新規手法を提案した．先行研究は，検出キーワード群に頼って区間を結合するか比較的短い文書を対象として単語ペアをモデリングするため，長い区間と短い区間の両方を正確に推定することが困難である．提案手法は，確率モデルに対象単語と文脈内各単語のペアに関するトピック毎の共起を導入し，新規に考案したマルコフ連鎖モンテカルロ法でパラメータ群の値を推定することにより，この問題に対処した．英語 Wikipedia 記事や英語携帯電話レビュー，人工データなどの文書データを用いた実験の結果，提案手法は，先端手法を含む比較手法をトピックの一貫性と区間分割の正確性で上回り，計算時間も比較的短いことが分かった．

最後に，入力された文書集合において，前段落に記した文脈を活用して，各出現単語が所属するトピック確率分布を推定する語意曖昧性解消問題に対して，多義語と類義語に関する語意も正確に推定する手法を提案した．先行研究は，単語クラスタリングを採用するか単語埋め込み法を併用するため，語意推定の正確性に難点がある．提案手法は，語意に関する確率変数を確率モデルに導入し，新規に考案したマルコフ連鎖モンテカルロ法でパラメータ群の値を推定することにより，この

問題に対処した．この確率変数は，対象となる単語出現に関する推定トピック確率分布と，同一単語の全出現に関する推定トピック確率分布の重みつき和として表される．英語ニュース記事，英語 Wikipedia 記事議論，英語ツイートなどの文書データを用いた実験の結果，提案手法は，先端手法を含む比較手法を，語意推定，文書分類，トピックモデリングの正確性で上回り，語意の変遷も検知できることが分かった．