

論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	Research on Context-based Document Topic Analysis
著者(和文)	呂有為
Author(English)	Youwei Lu
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第10664号, 授与年月日:2017年9月20日, 学位の種別:課程博士, 審査員:新田 克己,樺島 祥介,渡邊 澄夫,小野 功,高村 大也
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第10664号, Conferred date:2017/9/20, Degree Type:Course doctor, Examiner:,,,,,
学位種別(和文)	博士論文
Category(English)	Doctoral Thesis
種別(和文)	審査の要旨
Type(English)	Exam Summary

論文審査の要旨及び審査員

報告番号	甲第	号	学位申請者氏名	呂 有為
		氏 名	職 名	
論文審査	主査	新田 克己	教授	高村 大也
審査員	審査員	樺島 祥介	教授	
		渡邊 澄夫	教授	
		小野 功	准教授	

論文審査の要旨 (2000 字程度)

本論文は「Research on Context-based Document Topic Learning」と題し、英文 5 章より構成されている。

第 1 章「Introduction」では、文書からトピック抽出を行う手法として確率トピックモデルを取りあげ、その代表的なトピックモデル LDA (Latent Dirichlet Allocation) を紹介している。実際のデータの解析において、教師なし学習である LDA では分析されるトピックは人間の理解と予測に繋がりにくい点、単語の順序を無視する LDA では文書のローカル情報に基づくトピックを分析することができない点で問題であるとしている。この二つの問題点を解決するため、トピック抽出の新しい手法を提案し、その評価をするのが本論文の目的であるとしている。

第 2 章「Semi-supervised Latent Dirichlet Allocation」では、半教師付き学習 LDA (ssLDA) を提案している。半教師付き学習では、一部の文書につけられたラベルをトピックモデルの中の潜在的トピックと結びつけ、そのラベル情報を利用したモデルを学習する。ラベルの付け方は応用によって方式が異なる可能性があるが、ssLDA は「word-level topic assignment」と「Document-level topic assignment」という二種類のラベリングに対応できるとしている。評価実験をマルチラベルの認識問題で行い、他の認識手法である Nearest Neighbor と Transductive SVM (Transductive Support Vector Machine) と比較した結果、ssLDA はこれらの手法より「label-based F1 score」と「document-based F1 score」の評価基準で性能が良いことを確認している。

第 3 章「Modeling Content Structures of Domain-specific Texts with RUP-HDP-HSMM」では、文書中の位置による情報の偏りを想定しトピックの推論を行う RUP-HDP-HSMM (Recursive Uniform Partitioning Hierarchical Dirichlet Process Hidden Semi-Markov Model) を提案している。さらに、そのトピックの系列で文書構造を表すことにより、文書検索やマイニングタスクなどに対し有用な情報を提供できると主張している。RUP-HDP-HSMM と従来の時系列データをモデリングするアルゴリズム HDP-HSMM との違いは RUP の使用にある。RUP は単語の時系列データをいくつかのセグメントに分割する確率モデルであるが、時系列が長くてもセグメントの数の平均値が緩やかに増えていくだけであることをシミュレーションを通じて確認している。この性質を持つ RUP を HDP-HSMM に組み込むことによって、より正確に文書構造をモデリングすることが期待できるとしている。RUP-HDP-HSMM による文書構造のモデリングの性能を評価するために、原発再稼働に関する日本の社説をテキストデータとし、文書検索、分類とマイニングを含む三つの実験を行い、その結果を、滞留時間が Poisson 分布と仮定した HDP-HSMM、滞留時間が負二項分布と仮定した HDP-HSMM、bag-of-words、LDA と比較している。1 番目の文書検索の評価実験では、提案手法が一番良い性能を示し、文書のローカル情報に基づくトピック分析は対象データの特徴をより正確に捉えることを確認している。2 番目のテキスト分類実験では、提案手法により見つかった文書構造は SVM などの分類器に有用な特徴量を提供できることを示している。3 番目のテキストマイニングでは、文書の 2 クラス分類問題において、正の事例から文書構造に基づきその分類を解釈する部分テキストを各文書から探索しその質を評価することにより、対象データセットにおいて点数が高い部分テキストは分類を解釈することが容易になるという結果を示している。

第 4 章「WPM: An Alternative to RUP」では、第 3 章の RUP をさらに拡張した文書分割確率モデル WPM (Weibull Partition Model) を提案している。WPM はいくつかの「key position」を設け、その周り

のセグメントの分布を柔軟に推定することが特徴である。WPM の構築には Weibull 分布が利用されている。セグメントの始点は「key position」であれば、終点の分布は明確に指定されるが、「key position」でなければ、終点の分布は隣接の「key position」の終点分布を用いて構成する。WPM の応用の一つである文書構造モデル WPM-HDP-HSMM は WPM と HDP-HSMM を組み合わせるモデルである。人工合成データを用いて WPM-HDP-HSMM で隠れ状態を推定する評価実験を行い、WPM-HDP-HSMM は HDP-HSMM や、隠れ状態の遷移を抑えるように設計された Sticky HDP-HMM のいずれよりも早く収束し隠れ状態の識別精度が優れた結果が得られたとしている。

第 5 章「Conclusion and Future Work」では、本論文で提案された 3 つのモデルについて議論するとともに、今後の発展の可能性について述べている。

以上を要するに、本論文は文脈情報を考慮するトピック分析に向け、人間の理解と予測に繋がりやすいラベル情報を学習に利用できる ssLDA と文書のローカル情報から文書構造を発見する RUP-HDP-HSMM と、RUP-HDP-HSMM を拡張する WPM-HDP-HSMM を提案している。この成果により、従来の典型的確率トピックモデルより、さらに有用な情報を文書から抽出することができるため、自然言語処理やテキストマイニングなどの様々なタスクで応用することができ、工学上貢献するところが大きい。よって、本論文は博士（工学）の学位論文として十分な価値があるものと認められる。

注意：「論文審査の要旨及び審査員」は、東工大リサーチリポジトリ(T2R2)にてインターネット公表されますので、公表可能な範囲の内容で作成してください。