OUKA
The University of Osaka
Institutional Knowledge Archive

Title	ユーザの意図理解を目的とした文書データからの知識獲得に関する研究
Author(s)	藤本, 拓
Citation	大阪大学, 2014, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/34562
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

様式3

論 文 内 容 の 要 旨

氏 名 （ 藤本 拓 ）	
論文題名	ユーザの意図理解を目的とした文書データからの知識獲得に関する研究

論文内容の要旨

本論文は，ユーザの意図理解を目的とした，高精度な知識獲得手法を確立する．文書データからの知識の獲得は，その目的に応じて必要とされる分析技術が異なる．そこで本論文では，以下に示す二つの軸によって分析技術を分類する．第一の軸は，文書分析の粒度である．文書集合を対象した分析か単一の文書を対象とした分析かによって，分析結果の適用先が大きく異なる．第二の軸は，時系列を考慮した分析を行うか否かである．時系列を考慮して文書を分析することで，静的な分析では得られないリアルタイムのホットトピック変化や異常の検知といった知識を獲得可能である．

本論文ではまず，文書集合に対する非時系列分析技術として，ユーザのWeb閲覧ログを分析する文書の対象とし，トピックモデルによる，ユーザのWeb閲覧行動の高精度なモデル化手法を提案する．提案方式は，Webページの概念的な階層関係を表すWebディレクトリ辞書を導入する．これにより，本方式においてトピックモデルの単語に置き換えられるユーザのWeb閲覧行動を抽象化し，最上位概念の行動を抽出することで，高精度なモデル化を実現する．本論文では，7537人から得た4か月間のWeb閲覧ログを利用した大規模な実験により，提案方式の精度が既存手法を大きく上回ることを示す．

本論文では次に，文書集合に対する時系列分析技術として，時系列の文書集合の変遷を高精度にモデル化する手法を提案する．提案方式は，トピックモデルを拡張し，時系列の文書集合を高精度にモデル化する．具体的には，トピックモデルにより各時刻の文書集合をモデル化し，これらを時系列に平滑化することで，時間変化にロバストなモデルを生成する．この際，時系列フィルタであるパーティクルフィルタを適用することで，最適な平滑化係数を導出する．さらに，時々刻々と登場する新語に対応してモデルが扱う語彙集合を動的に拡張する．本論文では，9ヶ月間1日100万のツイートデータを利用した大規模な実験により，提案方式の精度が既存手法を上回ることを示す．

本論文ではさらに，単一文書に対する非時系列分析技術として，ユーザの発話を単一の文書とし，発話からユーザの意図を高精度に理解する手法を提案する．提案方式は，別途用意した文書集合から，単一文書の意図理解に適した学習モデルを生成し，これに従って文書を生成したユーザの意図理解を行う．この際，単一文書からのユーザ意図理解に有効と思われる特徴を導入することで，モデルの精度向上を図る．本論文では，3008人の被験者から収集した発話例を元に学習モデルを構築し，提案方式による意図理解の精度を評価し，高い精度でユーザの発話意図を理解可能であることを示す．

最後に結論として上記の研究を総括し，課題と今後の展望を示す．

様式7

論文審査の結果の要旨及び担当者

氏　名　（　藤本　拓　）			
論文審査担当者		（職）	氏　名
	主 査	教授	西尾 章治郎
	副 査	教授	藤原 融
	副 査	教授	細田 耕
	副 査	教授	薦田 憲久
	副 査	教授	下條 真司
	副 査	准教授	原 隆浩
	副 査	准教授	前川 卓也

論文審査の結果の要旨

ソーシャルメディアの発展に伴い、インターネット上には大量の文書データが蓄積されつつあり、これらのデータから様々な知識を獲得することが可能となった。特に、文書を生成したユーザの意図を理解することは、トレンド分析、コミュニティ分析、評判分析等、様々なサービスへの応用が可能であり、近年になって様々な技術が提案され注目を集めている。しかし、既存の技術は、いずれも文書データをモデル化するに当たって技術的な課題があり、十分高精度にユーザの意図を理解可能であるとは言えない。この課題に対し、本論文では、ユーザの意図を理解するための文書モデル化技術を、文書分析の粒度と時系列分析の有無という独自に定義した二つの軸によって分類し、各分類について、既存技術を上回る精度で文書データをモデル化可能な手法を提案し、さらに実アプリケーションを考慮した大規模な実験により、その有効性を示している。本論文の主要な研究成果を要約すると次の通りである。

(1) 第一の分類として、非時系列の文書集合を高精度にモデル化する手法を提案している。この手法では、木構造に構造化された単語間の概念辞書を構築し、トピックモデルの入力となる文書データに含まれる単語を概念辞書によって抽象化することで、文書集合を高精度にモデル化する。
(2) 第二の分類として、時系列の文書集合を高精度にモデル化する手法を提案している。この手法では、時系列のトピックモデルにおいて、パーティクルフィルタを利用することで時系列のモデル変化を最適にスムージングすると共に、新語を動的にモデルに追加することで、文書集合を高精度にモデル化する。
(3) 第三の分類として、非時系列の単一文書を高精度にモデル化する手法を提案している。この手法では、事前に収集した大量の学習データを利用して、文書を分類するモデルを生成する。この際、短い文書に効果的な特徴量を導入した上で、実験的に最適な組み合わせを探索することで、高精度なモデルの生成を行っている。

以上のように、本論文はユーザの意図理解を目的とした自然言語処理技術の構築における先駆的な研究として、情報科学に寄与するところが大きい。よって本論文は博士（情報科学）の学位論文として価値のあるものと認める。