確率モデルに基づく自然言語文書からの知識抽出に関する研究

白井, 匡人 / SHIRAI, Tadahito

(出版者 / Publisher)
法政大学大学院理工学・工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学・工学研究科編 / 法政大学大学院紀要. 理工学・工学研究科編
(巻 / Volume)
57
(開始ページ / Start Page)
1
(終了ページ / End Page)
7
(発行年 / Year)
2016-03-24

博士学位論文
論文内容の要旨および審査結果の要旨

論文題目	確率モデルに基づく自然言語文書からの知識抽出に関する研究
氏　　名	白井　匡人
学位の種類	博士（工学）
学位授与年月日	2016 年 3 月 24 日
学位授与の条件	法政大学学位規則第 5 条第 1 項第 1 号該当者（甲）
論文審査委員	主　　査　　三浦　孝夫　　教授
	副　　査　　斉藤　利通　　教授
	副　　査　　加藤　豊　　教授
	副　　査　　佐藤　哲司　　教授（筑波大学図書館情報メディア系）

2015 年 12 月 28 日

学位論文審査委員会

委員長　佐藤　勉　　殿

学位論文審査小委員会

主査　教授

副査　教授

副査　教授

副査　筑波大学 図書館情報メディア系

試問による学識確認の報告

法政大学学位規則第１９条により、　白井　匡人　氏について、その論文を中心に関連する学問領域の試問を行った結果、合格と判定した。

以　上

（報告様式Ｉ）

2015 年 12 月 28 日

学位論文審査委員会

委員長 佐藤 勉 殿

学位論文審査小委員会

主査 教授 三浦孝夫 印

副査 教授 斎藤利通 印

副査 教授 加藤尊雄 印

副査 筑波大学 図書館情報メディア系 教授

佐藤哲司 印

白井 匡人 氏 提出学位請求論文

「確率モデルに基づく自然言語文書からの知識抽出に関する研究」

論文内容の要旨と審査結果の要旨（報告）

（報告様式Ⅱ）

1．論文内容の要旨

近年，インターネットの発達から大量のデータを容易に入手できるようになっている．これらの多種多様なデータは，様々な情報を含むが定型化されていないため，未整理のまま流れ去っている．多くのデータはテキスト形式で記述されている．特にストリーム中の文書は，タイムリな情報を扱うことから，極めて重要な情報源として注目されている．この大量のデータから知識を抽出するために，自然言語文書の解析手法が必要となる．自然言語文書の解析は，過去の様々な研究で扱われてきたが，コンピュータによる本質的な文書の理解には至っていない．自然言語文書の解析では，多義語や同義語による文脈曖昧性・語義曖昧性の解消といった様々な問題が残されている．

文書の知識は，文書の持つ何らかの意味による記述を表す．定性的には，パターンやその組み合わせによる記述であり，言語モデルや論理モデルによって扱われる．定量的には，パラメータやパラメータモデルによる記述であり，統計モデルや確率モデルによって扱われる．また，文書の知識は，特定の知識を抽出するなら，知識獲得手法(決定木やベイズ分類)を用いる．抽出した意味が不明であるとき，潜在意味抽出やクラスタリングを用いる．

文書から抽出できる表層的な情報として，単語や品詞，フレーズがある．単語の出現品度や共起頻度は，文書の特徴となりえるが，高頻度語と重要語は異なる．熟語やコロケーションは一つのまとまりとして意味を成すため，個々の単語と区別する必要がある．また，表層的な情報は，語義や構文の曖昧さに強く影響を受ける．これにより，正確な意味を捉えることができない可能性がある．単語の字面を扱うだけでは構造を区別することができない．統計手法は表層的な情報しか扱えず，表現能力に限界がある．例えば，主成分分析や潜在意味分析を行うことにより，文書の持つ潜在的な意味を抽出する試みがなされている．しかし，抽出した語のまとまりを人手によって解釈する必要がある．

本研究の狙いは，確率モデルによる文書集合の高度なモデル化の試みである．確率モデルは，確率分布により，文書集合全体を表現できる．確率モデルの利点としては，事前分布を用いることで文書に出現しない語を扱え，変数の依存を条件付き確率で表現できることが挙げられる．また，文書の単語分布は，多項分布に従うことが経験的に知られている．

過去の多くの研究でも，確率モデルによる文書モデル化が行われている．
ユニグラムモデルは，確率モデルに基づく文書集合のモデル化である．ユニグラムモデルでは，全ての文書集合を一つの確率分布で表す．混合ユニグラムモデルでは，各クラスが確率分布を持つため，クラスごとの分布の異なりを表現できる．しかし，これらのモデルは文書が 1 つの確率分布に従うことを仮定する．このため，共通して表れる単語や，分野に依存してよく使用される単語といった，単語を出力する要因の混合を考慮することができない．本来単語が持つ意味は文書中の話題や文脈に依存して異なる．単語の意味の特定は，自然言語の持つ曖昧性によって定型化して抽出できないという点で困難である．自然言語文書から知識抽出を行うには，文書を特徴付ける要因の混合を考慮する必要がある．

本研究では，単語の意味や文書の内容を潜在状態を用いて確率的に捉え，単語の持つ意

味を考慮した解析を可能にする．これによりクラスや話題が混合した文書集合から高水準な知識抽出を行う．確率モデルは，与えられた文書集合を観察値と見て推定される．しかし，動的な文書集合では，扱われる話題が時間と共に変動するため，特徴の変化が頻繁に起きる．文書ストリーム（ニュース記事等の新たな文書が逐次発生する文書集合）では，特徴の変動とデータ量が大量となることが知識抽出を困難にする要因となる．文書を特徴付ける要因の変化を捉えることがストリーム中での知識抽出において重要となる．また，ストリーム中では，変化に応じたモデルの修正が必要である．

本学位請求論文では，潜在要因を考慮した確率モデルに基づき自然言語文書からの知識抽出を行う手法を提案し，主に 2 つの問題について論じた．すなわち，文書を特徴付ける複数の要因が混在した文書集合からの特徴抽出と特徴の変化が起きる文書ストリームからの特徴抽出である．第 1 の問題では，トピックモデルにより単語の持つ潜在トピックを推定し，文書を潜在トピックの混合で表した．この潜在トピックを文書を特徴付ける要因と見なすことで，検索語の意味の抽出といった自然言語文書からの知識抽出が行えることを示した．また，日本語文書の品詞分布がジャンルを特徴付ける要因となることを示した．第 2 の問題では，ストリーム中での文書の特徴を抽出するためにトピックモデルのオンライン学習を行った．文書ストリーム中での特徴の変動を事前分布の変化と対応させることで特徴抽出を行った．また，ストリーム中での特徴を定常分布と変動分布として抽出した．

本研究では，これらの問題を網羅的に扱い，各問題を解決するための手法を提案した．

第 1 章では，自然言語文書からの知識抽出の概要と課題について論じた．また，本研究で取り組む，文書を特徴付ける複数の要因が混在した文書集合からの特徴抽出と，文書ストリームからの特徴抽出に対する問題点について論じた．

第 2 章では，本研究の研究背景と課題について論じた．ここでは，各問題に関する先行研究について述べた．

第 3 章では，潜在トピックによる著者の特徴を抽出する手法について論じた．提案手法は，トピックモデルを用いて単語の潜在状態を推定することで文書を潜在トピックの混合で表現する．これにより，クラスや話題といった文書を特徴付ける要因が混在した文書集合において高精度に分類が行えることを示した．

第 4 章では，日本語の品詞分布特性によるジャンルの特徴を抽出する手法について論じた．ここでは，日本語文書の品詞分布の特性をジャンルごとに仮定することで，文書の品詞情報のみを用いてジャンル分類を行えることを示し，品詞分布が文書を特徴付けていることを示した．

第 5 章では，検索語の意味を抽出する手法について論じた．検索語間の依存関係を捉えることで，検索語の意味を考慮して文書検索を行えることを示した．検索語に対して係り受け語を推定し，その組に対してトピックを推定することで，検索語の意味を抽出した．

第6章では，事前分布の学習による動的な特徴を抽出する手法について論じた．ストリーム中でのクラスの出現確率やトピックの確率分布，トピックの単語分布の変化を事前分布の変化として捉えることで，変動する特徴の抽出を行った．

第7章では，オンライン学習によるストリーム中の特徴を抽出するについて論じた．ストリーム中でオンライン学習を行うことで，文書集合の特徴の変化を抽出した．ストリーム中での定常的な特徴を定常トピック，変動的な特徴を変動トピックを用いて表現することで文書ストリームの分類を行えることを示した．

第8章では，複数のラベルを持つ文書からの特徴を抽出する手法について論じた．ここでは，文書中のラベルの混合と話題の混合を同時に扱えるようにトピックモデルを拡張した.ストリーム中でラベル間の共起関係を抽出することで，複数のラベルを持つ文書の分類を行えることを示した．

第9章では，本研究の結論と応用について論じた．また，本研究の今後の発展について述べた．提案手法を用いることで各問題を解決し，自然言語文書からの知識抽出を行うことができることを述べた．

2．審査結果の要旨

本学位請求論文は，確率モデルにより自然言語文書からの知識抽出を行い，その有効性を示したものである．審査の結果，以下の点で工学的新規性と有効性を確認した．

(1) 潜在要因による特徴抽出

確率モデルに潜在要因を仮定することで自然言語文書の特徴を表現する手法を提案した．ここでは，トピックモデルにより単語の持つ潜在トピックを推定し，文書を潜在トピックの混合で表した．この潜在トピックが文書を特徴付ける要因と見なすことで，検索語の意味の抽出等の自然言語文書からの知識抽出が行えることを示した．また，日本語文書の品詞分布がジャンルを特徴付ける要因となることを示した．

(2) 文書ストリームからの特徴抽出

確率モデルの更新により文書ストリーム中での特徴の変化を捉える手法を提案した．ここでは，トピックモデルのオンライン学習により，ストリーム中の特徴の変化を表現できることを示した．文書ストリーム中での特徴の変動を事前分布の変化と対応させることで特徴抽出を行った．また，ストリーム中の文書から定常トピックと変動トピックを抽出することで定常的な特徴と変動的な特徴を表せることを示した．

すなわち，本論文は学術的な観点から新規性に優れており，工学的応用の基礎としての意義も認められる．よって，本審査小委員会は全会一致をもって提出論文が博士(工学)の学位に値するという結論に達した．