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

Studies on Knowledge Extraction based on Probabilistic Model from Natural Language Documents

白井 匡人♡

Masato SHIRAI

本論文では，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る．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す．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す．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う．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表現する．また，ストリーム中の文書の特徴の変化を捉えるために定常的な特徴と変動的な特徴をそれぞれ定常トピック，変動トピックとして抽出する．

1 1. 前書き

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，文脈の理解や談話の理解，多義語・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．単語の出現頻度や共起頻度は，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，統計能力に限界がある．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に対する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行う．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．

本論文は，自然言語文書からの知識抽出を行うために 2 つの問題を論じる．第 1 の問題は，クラスや話題といった文書を特徴付ける要因の抽出である．文書の単語の分布には書き手やテーマ，ジャンルなど複数の要因による影響が混在していると仮定する．ここでは確率モデルを使用し，各単語の潜在状態を推定する．具体的には品詞分布の法則や潜在トピックを扱う．第 2 の問題は，動的な文書集合からの知識抽出である．ここでは，ストリーム中での各要因の特徴の変化が潜在トピックの変化として表現できることを仮定する．文書集合の変化に対応するため，ある間隔で到着する教師データを用いて新たに学習することにより，モデルを修正する．また，ストリーム中の文書には大きく 2 つの特徴がある．時間に影響されず各文書内で共通して表れる定常的な特徴と，文書中の話題の変化による特徴の変動である．このため，文書集合の特徴を表す分布は，1 つの定常状態に収束するのではなく，時間とともに変化することを仮定する．提案手法はオンライン学習を用いて特徴の変化に応じて潜在要因の変化を学習する．ここでは，事前分布の変化の学習と定常分布と変動分布の抽出を行う．本論文の論点は主に，(1) トピックモデルの適用範囲の拡大 (2) 文書ストリームへのトピックモデルの適用，という 2 点にある．

2. 潜在要因による特徴抽出

本論文は，文書を特徴付ける要因として単語の潜在要因に着目する．ここでは，潜在要因として潜在トピックと品詞を用いる．この潜在要因は確率モデルにより抽出する．トピックモデルは，文書を潜在トピックの混合で表す確率モデルである．一般的に，トピックモデルは多重集合 (bag of words) 表現された文書を対象とする．多重集合は，文書中の単語の並びを考慮せず，各単語の頻度 $v_1, v_2, \cdots, v_n$ のみを表現する．

♡ 正会員　法政大学理工学研究科
masato.shirai.2d@stu.hosei.ac.jp

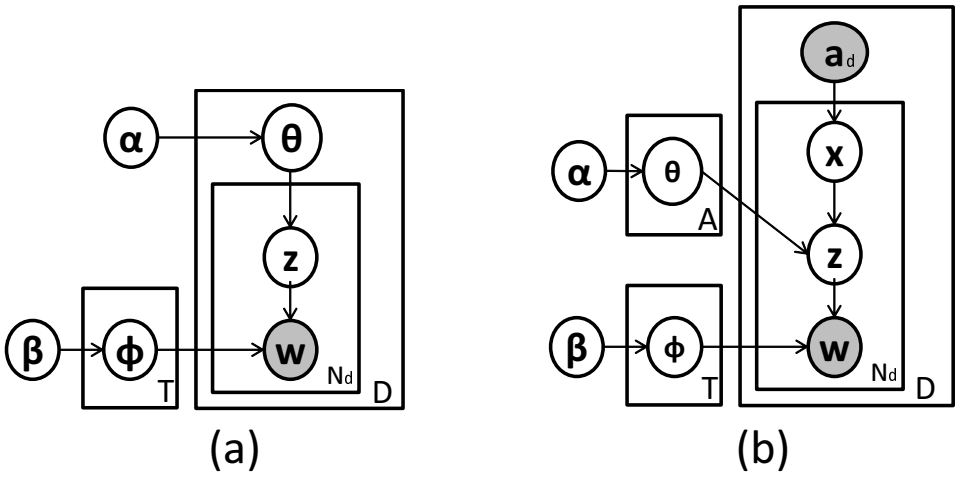

図 1: LDA と AT モデルのグラフィカルモデル

2.1 潜在トピックによる著者の特徴抽出

本研究では，文書が持つ複数の要因を抽出するために，トピックモデルにより著者の特徴を抽出する．文書はクラスが持つ複数の話題を含み，話題に関連して出現する単語分布が変化する．ここでの問題は，小説や社説などの文書には著者の特徴が表れるが，単語の分布の類似が著者によるものか，同一の話題によるものか判別することができない点にある．SF やミステリー等，同じジャンルを扱っていればジャンルに依存した単語が出現することが考えられる．同じ話題を扱っている文書には，共通の固有名詞などが多数使用される．このため，複数の要因を含む文書を扱うには，文書内の話題の混合を表現するモデル化が必要不可欠になる．提案手法は，トピックモデルを用いることで，文書を特徴付ける要因を捉える．トピックモデルでは，著者や新聞社といったクラスの特徴を話題の混合として表現する．

LDA(Latent Dirichlet Allocation) はトピックモデルの一種であり，1 つの文書が複数の潜在トピックの混合として表現される [1]．トピックモデルでは，文書内の各単語に潜在状態であるトピックを仮定し，単語に潜在トピックを 1 対 1 で割り当てる．Author-Topic(AT) モデルは著者ごとにトピック分布を持つ点で異なる [3]．AT モデルは複数の著者によって書かれた文書を扱えるが，本研究では単純化のため，各文書が単一の著者によって書かれている文書を扱う．このトピックモデルを用いることでクラスと各文書を潜在トピックの混合として表すことが可能となる．潜在トピックは，共通したクラスや文書に出現するといった何らかのまとまりを示す一種のクラスタである．潜在トピックが話題に対応すると仮定することで，話題の変化を潜在トピックの変化として表現する．しかし，小説や社説といった複数のジャンルを含む文書集合においても，適切なトピックを学習できるのか定かではない．

図 1(a) では，LDA のグラフィカルモデルを示す．図中の変数は，図 1 左下に，ディリクレ事前分布 $Dir(\beta)$，図 1 左下の単語空間の多項分布 $Multinomial(\phi_{z_i})$, T はトピック数，図 1 左上にディリクレ事前分布 $Dir(\alpha)$，図 1 中央にトピック空間の多項分布 $Multinomial(\theta_d)$, D は文書数，N_d は各文書の単語数を表す．LDA の単語生成過程を以下で示す．まず，すべてのトピック t においてディリクレ事前分布 $Dir(\beta)$ から ϕ_t を抽出し，同様に，すべての文書 d においてもディリクレ事前分布 $Dir(\alpha)$ から θ_d を抽出する．次に，文書 d 内の i 番目の単語 w_i において，抽出した文書 d の多項分布 $Multinomial(\theta_d)$ からトピック z_i を抽出し，そのトピック z_i の多項分布 $Multinomial(\phi_{z_i})$ から単語 w_i を抽出する．

LDA の拡張である AT モデルでは，図 1(b) で表される変数 x により文書の著者などの該当の文書が属するラベルの情報を持つ [3]．文書 d において $\mathbf{a}_d$ から著者 x は一様に得られ，トピック z は文書のトピック分布に基づきトピックから単語が生成される．LDA は文書ごとにトピック分布を持つが，AT モデルでは著者ごとにトピック分布を持つ点で異なる．また，AT モデルは文書に複数のラベルが存在することを仮定している．しかし，文書内でのラベルの偏りを考慮しておらず各ラベルは一様に生成される．

本研究では，トピックモデルを用いて各クラスを潜在トピックの確率分布として表現する．KL 情報量やカイ 2 乗値により各クラスとテスト文書のトピック分布を比較することで著者推定を行う．ここでトピック分布は小説と社説が混在した文書集合から求める．これにより，異なるジャンルの文書が混在した文書集合においても，著者ごとの分布の異なりを正しく捉えることを示す．

2.2 日本語の品詞分布特性によるジャンルの特徴抽出

文書が持つ複数の要因の抽出を行うために，ジャンルを特徴付ける要因を抽出する．文書ジャンルとは，文書の役割・状況を考慮した文書の種類であり，典型的には ”日記”, ”随筆”, ”小説”, ”社説”, ”報道記事”などがある．ここでは，ジャンルを特徴付ける要因として品詞に着目する．自然言語文書は，何らかの文法規則に基づき記述される．品詞は単語の働きに応じて種類分けされたものであり，文の意味を解釈するうえで有用である．

日本語文書の品詞分布に関しては，いくつかの先行研究によって品詞分布の法則が示されている [6][4]．樺島の研究では，名詞の割合によって他の品詞の割合が定まるとし，樺島の近似式が示されている．大野らは，延べ語数，見出し語数のどちらにおいても日本語文書には品詞分布に法則性が存在することを指摘している．しかし，これらの先行研究で示されている近似式は品詞分布の特性を表してはいるが，実際の文書に適用すると当てはまりがあまり良くない．この原因としては，品詞の割合には文書のジャンルによって差がある [5] としながら，ジャンルごとの品詞分布の違いを考慮していないためであると考えられる．このため，品詞分布の法則がジャンルごとに成り立つと仮定する．品詞分布の法則がジャンルごとに適用可能ならば，品詞の割合だけからジャンル推定でき，極めて効率よく分類できる．

提案手法は，図 2 のようにジャンルごとに名詞の割合が変化すると仮定し，名詞の割合の事前分布にガウス分布を用いる．

2.3 検索語の意味抽出

文書を特徴付ける要因として，検索における語の働きを論じる．多クラス分類では，クラスを集合として与え，その特徴を基に各クラスに該当する文書を抽出する．これに対し，文書検索は

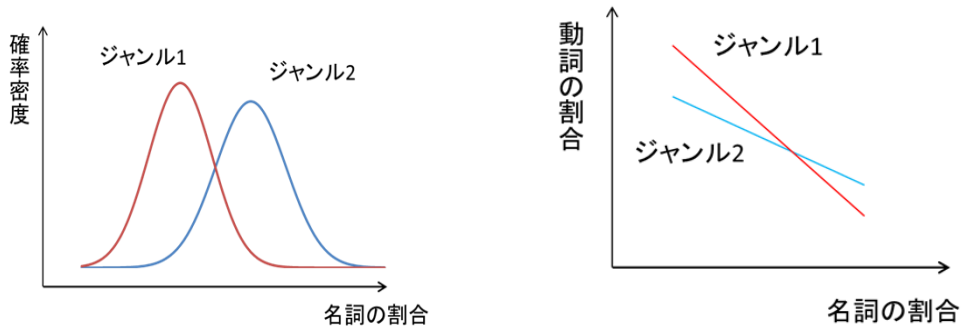

図 2: ガウス分布と線形回帰

数個の単語からなる検索語に対応する文書を抽出する．検索語は一般的に数個の単語からなり，この検索語との類似性により関連する文書を選択する．ここでの問題は，検索語の語義の曖昧性により，質問者の要望と異なる文書が選択される点にある．例えば，”APPLE”,”PIE”,”CHART”という 3 語からなる検索語から，質問者がアップルパイの売れ行きを知りたいのか，アップルコンピュータに関する円グラフを調べたいのかということを判断することは難しい．このため，検索語全体の意味を考慮して抽出する必要がある．

提案手法は，単語の係り受け関係を用いることで単語の依存構造を抽出する．さらに単語と係り受け語の組に対して潜在トピックを推定することで検索語の意味を考慮した検索を行う．

3. 文書ストリームからの特徴抽出

本研究の目的は，潜在要因の変動を考慮することにより，動的な文書集合から特徴抽出を行うことにある．ニュースストリーム，マイクロブログに代表される SNS データは，極めて重要な情報源として注目されており，多数の人々が利用している．文書ストリームでは，一度新たな話題が発生すると関連した記事が短期間に多数発生し，単語分布が大きく変わるというバースト現象が発生する [2]．各話題に依存した単語分布は，新たなイベントの影響を強く受ける．例えば，オリンピックのような大きなイベントが起これば，そのイベントに注目が集まるため期間中はオリンピックに関する記事が急増する．期間が終わると異なる話題に遷移していきオリンピックに関する記事は減少する．

3.1 事前分布の学習による動的な特徴抽出

本研究では，文書ストリームからの特徴抽出を行うために事前分布の更新を行う．文書はクラスが持つ複数の話題を含み，話題に関連して出現する単語分布が変化する．ストリーム中では，この話題の特徴が大きく変化する可能性がある．この話題の変化によってクラスの特徴が変化していくため，動的な学習によりモデルを更新する必要がある．提案手法は，この特徴の変化を潜在トピックや単語を出力する事前分布の変化として捉える．

提案手法は，トピックモデルのオンライン学習により各クラスについてのディリクレ事前分布と各クラスの文書の特徴を学習する．トピックモデルのオンライン学習により事前分布を更新することで，クラス内での分布の変動を表す．バースト現象はクラスごとに生じ，変動が他クラスに影響を与えない．どのクラスでバーストが生じるかは多項分布確率的に生じると仮定する．従って複数のトピックモデルを混合して協調する上位モデルが必要である．クラスの出現確率とトピックの確率分布，単語の確率分布の事前分布をそれぞれ学習することで動的な文書集合からの特徴抽出を行う．

3.2 オンライン学習によるストリーム中の特徴抽出

文書ストリームからの特徴抽出を行うためにトピックモデルのオンライン学習を行う．これにより，ストリーム中の文書から定常的な特徴と変動的な特徴を抽出する．クラスを表す定常的な特徴は，クラス内の話題によらず一定して表れる．ストリーム中の話題の変化による特徴の変動は，時期やイベントに依存することから，現在の状態に合わせて特徴を学習する．この 2 つの特徴は，異なる性質を持つことから，それぞれ別の確率分布として用いることで高精度に文書集合をモデル化できる可能性がある．

提案モデルは，文書のクラスラベルを学習データとして使用することで，クラスごとのトピック分布を学習する．定常トピックと変動トピックという 2 種類の潜在トピックを用いることで，ストリーム中での潜在トピックの特徴の変化を捉える．定常トピック分布は，これまでに出現したすべて文書から得られるクラスの特徴を表し，変動トピック分布はウィンドウにより直近の文書の特徴を表す．ストリーム中では特徴の変動が不定期に起きる可能性があるため，直近の文書に対してウィンドウを遷移させていくことで特徴の変化を考慮する．

3.3 ストリーム中の複数のラベルを持つ文書からの特徴抽出

文書ストリーム中で複数ラベルを持つ文書から特徴抽出を行うためにトピックモデルを適用する．これにより，複数のラベルを持つマルチラベル文書から各ラベルの特徴を抽出する．マルチラベル文書中には話題の混合とラベルの混合による特徴が同時に出現する．このため，文書を特徴付ける要因を抽出することがより困難になる．また，ラベル間には”経済”と”市場”は共起しやすいが”芸術”と”健康”は共起しにくいといった依存性があることが考えられる．マルチラベル文書を扱うには，ラベル間の関係を考慮することが重要となる．文書を特徴付ける要因の混合と合わせてラベルの混合を扱うためのモデルの拡張が必要となる．

提案手法は，トピックモデルを用いることでマルチラベル文書をモデル化する．ここでは，ラベルの定常的な特徴とストリーム中に発生する局所的な変動を考慮する．各ラベルの特徴を学習し，ラベル間の共起関係を用いることで特徴抽出を行う．

4. 結論

本研究では，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た．自然言語文書からの知識抽出では，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．

[文献]

[1] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, Journal of Machine Learning Research 3, pp.993-1022, 2003

[2] Kleinberg, J.: Bursty and Hierarchical Structure in Streams, Proc. 8th SIGKDD, pp.91-101, 2002

[3] Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents UAI '04 Proceedings of the 20th conference on Uncertainty in artificial intelligence, pp.487-494, 2004.

[4] 大野 晋: 基本語彙に関する二三の研究国語学,vol24,pp.34-46,1956

[5] 樺島 忠夫: 現代文における品詞の比率とその増減の要因について国語学,vol18,pp.15-20,1954

[6] 樺島 忠夫: 類別した品詞の比率に見られる規則性国語国文,vol250,pp.385-387,1955

白井 匡人 Masato SHIRAI
法政大学理工学研究科