自然言語文書の分類に関する研究

白井, 匡人 / SHIRAI, Masato

(発行年 / Year)
2013-03-24
(学位授与年月日 / Date of Granted)
2013-03-24
(学位名 / Degree Name)
修士(工学)
(学位授与機関 / Degree Grantor)
法政大学 (Hosei University)

2012年度　修 士 論 文

自然言語文書の分類に関する研究

STUDIES ON DOCUMENT CLASSIFICATION FROM NATURAL LANGUAGE DOCUMENTS

指導教官　三浦 孝夫 教授

法政大学大学院工学研究科
電気工学専攻修士課程

11R3122　白井 匡人
Masato SHIRAI

目 次

第1章　序論

1.1　問題の背景

近年，インターネットや計算機の普及により大量の文書が容易に入手できるようになっている．これらの文書には多種多様な情報が含まれているが，未整理の大量の文書を人手で分類することは困難である．このため自動的に分類を行う手法が求められている．計量文体分析の分野では一世紀以上も前から，文章を構成する要素の特徴を定量的に分析する方法で文章の執筆者の推定などを行っている．このような手法は，手紙や脅迫文の著者同定，スパムメール識別，ブログ分類にも応用されている．一般的は応用分野として自然言語文書の分類を行うことで，アンケート分析などを用いた顧客サービスの自動評価，テキストマネジメントや情報統合などに役立てることが可能となる．

文書分類では文書から分類の対象となる著者やジャンルなどのクラスの特徴を抽出し，クラスが未知の文書と各クラスの特徴を比較することで行われる．文書を構成する文章テキストは何らかの規則に従っているが定型化されていないため，単語やフレーズ，品詞など何らかの単位に分割し，それらの出現頻度や共起関係などを抽出して解析を行う．ここで未知の文書は特徴が最も似ているクラスに分類されるため，クラスごとに異なって特徴付けられていないと分類が行えない．このため分類行う際には，文書から特徴を抽出するだけでなく，分類の基準となるクラス独自の特徴を抽出する必要がある．

文書の著者を推定する著者推定では，文章から著者の持つ単語分布を推定して分類に用いる．同様に文書を小説や日記などのジャンルに分類するジャンル分類では，ジャンルが持つ単語分布の推定を行う．しかしながら，文章で使われる単語は書き手やテーマ，ジャンルなど複数の要因による影響が混在しているため，単語分布の偏りがどの要因によるものか定かではない．このため単語分布を直接著者やジャンルに対応付けることができないという問題がある．有効な分類を行うには単語分布以外の特徴を抽出することが必要となる．また，ニュース記事などのストリームデータでは新たな文書が次々と到着することから文書集合が逐次変化する．動的に変化する文書集合に対して分類を行うには分類の対象となるクラスの特徴の変化に応じて分類基準を動的に変更する必要がある．さらには，文書集合から何らかの話題に対する人の繋がりを抽出するコミュニティ抽出では，著者と話題によって分類を行うため，著者の特徴と話題の特徴の両方を捉えて分類を行う必要がある．高精度な分類を行うには，様々な要因が混合する単語分布からの著者やジャンルなどのクラスの特徴の推定，動的に変化

する文書集合の特徴の変化に応じて分類基準の変更，著者と話題を複合してクラスの特徴を推定することが必要となる．

本研究では，これらの問題を解決するために，トピックモデルと品詞分布を用いた分類法を提案する．これにより著者やジャンル，新たな文書が次々に到着する文書ストリーム，著者と話題の複合であるコミュニティの分類を行えることを示す．

1.2 扱う問題

1.2.1 トピックモデルによる著者推定

小説や社説などの文書には著者の特徴が表れていると考えられるが，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は困難である．LDA(Latent Dirichlet Allocation) はトピックモデルの一種であり，1 つの文書が複数のトピックの混合として表現されるという仮定である． LDA ではトピックと単語の多項分布にそれぞれディリクレ事前分布を導入し，混合比を事前分布から動的に生成する． LDA は学習データだけに依存しないが， 代わりに特定の確率モデルを仮定するため， 柔軟な観点で文書を扱える可能性がある．

本研究では，トピック分布を小説の著者や新聞社に対応させることで著者推定を行い，複数ジャンルの場合でも著者ごとの分布の異なりを正しく捉えることを示す．

1.2.2 品詞分布によるジャンル分類

トピックモデルによる文書分類は非常に高い精度で分類が行えるが，学習時間が膨大になるという問題があるため大規模な文書分類には不向きである．

そこで高速な分類を行うために文書の特徴量として品詞分布を用いる．品詞分布を特徴量として用いてジャンル分類が高精度に実施可能ならば，次元数が品詞の種類数となり，小さい次元で文書の比較を行うことができる．この結果，次元数が極めて少ないことでデータ数が増えても必要なメモリ量は少量で済むという利点がある．古典的計量言語学では日本語文書の品詞の分布に関して，名詞以外の品詞の割合は名詞の割合によって一意に定まるとし近似式が示されている．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，ジャンルごとの品詞分布の違いを考慮していない．

本研究では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たジャンル分類を提案する．これにより品詞分布の特性を活かし，ジャンル類別が行えることを示す．

1.2.3　品詞分布による文書ストリームの分類

ストリームデータは新しいデータが逐次到着することからデータ量が膨大となる．文書集合が動的に変化するため分類基準も動的に更新する必要がある．

日本語文書の品詞分布には法則性があることが知られており，この法則性を用いることでジャンル分類が行えることが示されている．特徴量として品詞分布を用いることで，ストリームデータのように各ジャンルの特徴が逐次変化する場合でもパラメータの更新が容易に行える．

本研究では，ガウス分布のパラメータと回帰直線の多項式係数ベクトルを差分学習することにより分類基準を逐次更新することで文書ストリームの分類を行う．

1.2.4　トピックモデルのラベリング

トピックモデルは目的に応じて潜在変数を加えることで柔軟にモデル化が行える．例えば文書に著者とコミュニティの潜在変数を与えることで Twitter に代表されるソーシャルネットワークからコミュニティを抽出できる．トピックモデルは対象の文書集合に応じて状態をモデル化できる手法として最も優れている．しかしながら，ここで著者とコミュニティはトピックによって特徴付けられているが，トピックとは確率的基準によって集められた一種のクラスタであり，潜在的トピックの混合で表される著者やコミュニティの意味は自明ではない．

本研究では，候補ラベルを用いて潜在的トピックをラベル付けし，トピック分布と著者分布を加味してコミュニティにラベル付けすることで，潜在的コミュニティに明示的な意味を付与する手法を提案する．

1.3　発表論文

1. 白井 匡人, 三浦 孝夫,”LDA を用いた著者推定”, 第 3 回データ工学と情報マネジメントに関するフォーラム (DEIM), 静岡, 平成 22 年 (2011) 2 月
   潜在的ディリクレ配分法 (LDA) では、１つの単語に付き１つのトピックを持ち、文書はトピックの確率分布で表される。同一の著者が書いた複数の文書を１つの文書と見なすことで、著者をトピックの確立分布で表せると考える。このトピックの確立分布を用いて著者推定を行う。また、実験によりその有効性を示す。

2. 白井 匡人, 三浦 孝夫,”On Domain Independence of Author Identification”, 12th Intn’l Conf. on Intelligent Data Engineering and Automated Learning (IDEAL), Norwich, 英国 , 平成 23 年 (2011) 9 月
   潜在的ディリクレ配分法 (LDA) では、１つの単語に付き１つのトピックを持ち、文書はトピックの確率分布で表される。同一の著者が書いた複数の文書を１つの

文書と見なすことで、著者をトピックの確立分布で表せると考える。このトピックの確立分布を用いて著者推定を行う。また、実験によりその有効性を示す。

3. 白井 匡人, 三浦 孝夫,"確率モデルによる品詞分布の特徴推定", データ工学と情報マネジメントに関するフォーラム (DEIM) , 神戸, 平成 24 年 (2012) 3 月
日本語文書の名詞に対する他の品詞の割合には相関関係があることが知られており、いくつかの近似式が示されている。しかしながら、すべての文書集合に同一の法則を適用する従来の手法では精度が悪い。そこで文書集合ごとにクラス分類を行うことで複数の近似式を得る。また、実験により本手法の有効性を示す。

4. 白井 匡人, 三浦 孝夫,"Document Classification Using POS Distribution", 16th East European Conference on Advances in Databases and Information Systems (ADBIS 2012), Poznan, Poland, 平成 24 年 (2012) 9 月
日本語文書の名詞に対する他の品詞の割合には相関関係があることが知られており、いくつかの近似式が示されている。しかしながら、すべての文書集合に同一の法則を適用する従来の手法では精度が悪い。そこで文書集合ごとにクラス分類を行うことで複数の近似式を得る。また、実験により本手法の有効性を示す。

5. 白井 匡人, 島田 諭, 三浦 孝夫,"品詞分布を用いた日本語文書のジャンル分類", 第 11 回 FIT（情報科学技術フォーラム）, 東京, 平成 24 年 (2012) 9 月
日本語文書の名詞に対する他の品詞の割合には相関関係があることを利用し、時系列的に並んだ文書ストリームをクラス分類する手法を提案する。ストリームデータは標本化により近似し、統計的に正しい範囲で分類基準を随時訂正する手法を示す。また、実験により本手法の有効性を示す。

6. 白井　匡人, 島田 諭, 三浦 孝夫,"トピックモデルのラベリングによる潜在的コミュニティの分析", 第 5 回データ工学と情報マネジメントに関するフォーラム (DEIM), 福島/郡山, 平成 25 年 (2013) 3 月
近年、トピックモデルを用いた研究が多方面で提案されている。例えば、ソーシャルネットワークのユーザ分析やコミュニティ抽出に用いる。しかしながら、ここで言うトピックとは、確率的基準による一種のクラスタであり、ユーザやコミュニティの意味を明示的に捉えることは自明ではない。本研究では、予め与えたラベルを用いて潜在的トピックをラベルつけし、明示的な意味を付与する手法を提案する。

第2章　LDAを用いた著者推定

潜在的ディリクレ割り当て (LDA) では，1つの単語に付き1つのトピック分布を持ち，文書はトピックの確率分布で表される．同一の著者が書いた複数の文書を1つの文書と見なすことで，著者を推定できる可能性がある．実際トピックの確率分布を用いテスト文書のトピックの確率分布と比較することで，著者推定を行うことができることを示す．

2.1　前書き

近年，文書の特徴を抽出するために様々な手法が用いられ，文書の特徴によって著者推定が行われている．著者推定に用いられる手法・特徴としては，出現する語のなんらかの特徴を捉えることが一般的である．構文解析後, 文節では名詞句など一般的意味を成す語は内容語，文節内で助詞など文法的な役割を果たす語は機能語と呼ぶ．機能語に関しては一語当たりの平均文字数 [5], 読点による特徴づけ [6]，n-gram 分布 [7] などが挙げられる．内容語の分布は著者の特徴となるが，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は困難である．

本研究では潜在的ディレクレ割当て (Latent Dirichlet Allocation, 以下 LDA)[1] を使うことにより，単語の潜在的トピックを推定する．これにより，文書を複数のトピック分布で表し，トピックの確率分布により著者推定を行う．LDA では，文書が1つのトピックに対応するのではなく，文書内に出現する各単語にトピックが確率的に対応すると見なす．この結果，各文書に複数のトピック分布が対応する．逆に，トピックの確率分布は文書ごとに存在するが，複数トピックをまとめて1つの文書に対応させ，著者にトピックの確率分布が対応すると考える．この仮説を著者トピックモデル (Author-Topic モデル)[9] という．著者が持つトピックの確率分布は著者の特徴を表したものである．文書が持つトピックの確率分布にはその文書の著者の特徴が表れる．LDA を用いれば，著者の単語の分布とテスト文書の単語の分布を比較することにより， トピックの確率分布によって著者推定が行える．本研究では，著者のトピックの確率分布とテスト文書のトピックの確率分布を (KL 情報量やカイ２乗値等を用いて) 比較する．

小説では SF やミステリー等，ジャンルは同じであっても他の著者と厳密に同じ話題について書かれている作品はほとんどない．社説のような他の新聞社と同じ話題について論じられることが多い文書を用いることによって，例え話題が同じであっても新聞社ごとに表れる新聞社の特徴によって著者推定が行えることを示す．著者推定は著

者の特徴によって推定を行うため，著者ごとに特徴があることが前提であり，著者の特徴量によって推定精度が変化すると考えられる．著者同士のトピックの確率分布を比較することで，他の著者との分布の異なりが大きい著者の推定精度が高くなると考えられるため実験で確認する．

第 2 章では LDA について述べ，第 3 章では推定手法について述べる．第 4 章では実験により有効性を示し，第 5 章で結論とする．

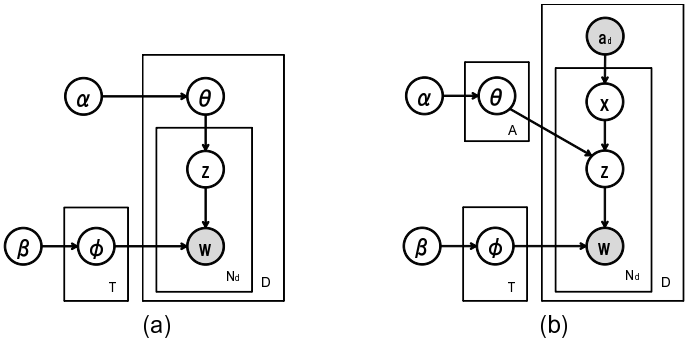

図 2.1: LDA と著者トピックモデルのグラフィカルモデル

2.2 LDA による著者推定

2.2.1 トピックモデル

トピックモデルとは, 1 つの文書が複数のトピックの混合として表現されるという仮定である．1 つの文書が 1 つのトピックで表される混合多項分布に比べ, トピックモデルは文書が複数のトピックの混合分布として, 各トピックが単語の分布として表現され, 高い精度で文書をモデル化する可能性がある. その中でも最近用いられているのが LDA である. 確率的潜在意味索引付け (Probablisitic Latent Semantic Indexing, pLSI) は, LDA と違って, トピックと単語の多項分布にそれぞれディリクレ事前分布を導入し, 混合比を学習データの文書集合に依存して固定化している.

一方, LDA ではこの混合比は事前分布から動的に生成する点で異なる．LDA は学習データだけに依存しないが, 代わりに特定の確率モデルを仮定するため, 柔軟な観点で文書を扱える可能性がある.

図 2.1 は LDA のグラフィカルモデルである. 図中の変数は, 図 2.1 左に, ディリクレ事前分布 $Dir(\beta)$, 図 2.1 左下の単語空間の多項分布 $Multinomial(\phi_{z_i})$, T はトピック数, 図 2.1 上にディリクレ事前分布 $Dir(\alpha)$, 図 2.1 中央にトピック空間の多項分布 $Multinomial(\theta_d)$, D は文書数, N 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.

2.2.2 著者トピックモデル

トピックモデルでは，著者は各文書にある様々な潜在的トピックについていくつかの文書を書くため，明示的な著者の情報は与えられない．そこで，トピックモデルに類似した著者トピックモデルでは，図 1 (b) で表される新しい変数 x により著者の情報を持つ．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確率的に選択され, 選択されたトピックから単語が生成される．ここで各文書はディリクレ事前分布から得られるトピックの分布と関連していると仮定する．このため，潜在的トピック z の多項分布から得られる単語 w はトピックについてのディリクレ分布 $p(\phi)$ から得られるトピック z と関連している．

著者トピックモデルでは文書が複数の著者によって書かれることを考慮しており，トピックモデルは著者が一人のときの著者トピックモデルと見なすことができる．

2.2.3 パープレキシティ

まず, モデルの評価として, モデル内での Topic, 繰り返し回数を決定する. そこで言語モデルの性能評価には一般に, テストセット・パープレキシティと呼ばれる情報理論に基づく客観的評価手法を用いる．テストセット・パープレキシティは次式で計算される.

$$PP = 2^{H(p)}, H(p) = -\frac{\sum_X \log_2 p(X)}{N}$$

ここで $w_1 \cdots w_N$ は評価用のテキスト集合とする． 各語の確率をもとめ，そのすべての情報量の平均をＥＸＰの係数としているすべての語の生起確率を求めるには，トピックの確率×語の生起確率を用いる．P_M は言語モデル M による $w_1 \cdots w_N$ の生成確率を表す. パープレキシティ値が高いほど単語の特定が難しく, 言語として複雑である. よって, パープレキシティの値が低いほど言語モデルの性能が高いと評価できる.

$$P_M(X_1 \cdots X_N) = \Pi_{i=1}^{N} p(X_i)$$

$$PP(s) = P_M(X_1 \cdots X_N)^{-\frac{1}{N}} = 1/\sqrt[N]{\Pi_i p(X_1) \cdots p(X_N)}$$

2.2.4 LDA のパラメタ推定

LDA のパラメタを推定する方法としては，EM アルゴリズムやギブスサンプリング等が挙げられるが，EM アルゴリズムでは局所最適解に陥りやすいという欠点がある

ため，本研究ではパラメタ推定にギブスサンプリングを用いる．ギブスサンプリングを用いてトピックの確率分布を更新することによって，各単語につくトピックが変化する．トピック j の確率分布は，トピック j 以外のすべてのトピックの確率分布によって更新される．これをすべてのトピックに対して行い，更新を繰り返すことにより，尤もらしい ϕ と θ の値が推定される．ある文書内では，ディリクレ分布によってトピックの確率分布には偏りがでるため，トピック内には，同じ文書で出現する単語が集まりやすくなっている．ここでギブスサンプリングの更新式を以下の式で定義する．

$$P(z_i = j | w_i = m, \mathbf{z}_{-i}, \mathbf{w}_{-i}) \propto \frac{C_{mj}^{WT} + \beta}{\sum_m C_{mj}^{WT} + V\beta} \frac{C_{dj}^{DT} + \alpha}{\sum_j C_{dj}^{DT} + T\alpha} \tag{2.1}$$

また，ギブスサンプリングの結果，推定される ϕ と θ の値は以下の式で求める．

$$\phi_{mj} = \frac{C_{mj}^{WT} + \beta}{\sum_m C_{mj}^{WT} + V\beta} \tag{2.2}$$

$$\theta_{dj} = \frac{C_{dj}^{DT} + \alpha}{\sum_j C_{dj}^{DT} + T\alpha} \tag{2.3}$$

C_{mj}^{WT} は単語 m がトピック j に割り当てられた回数，C_{dj}^{DT} は文書 d がトピック j に割り当てられた回数，V は全単語数，T は全トピック数である．

2.3 著者推定

著者が既知であり同じ著者が書いた文書を 1 つの文書と見なした文書の集合を Y，著者が未知である文書の集合を Z とする．この文書の集合である Y と Z を同時に LDA にかけることによって，著者ごとのトピックの確率分布 θ とテスト文書のトピックの確率分布 θ とトピックごとの語の確率分布 ϕ を得る．ここで得られたトピックの確率分布 θ は，各文書ごとに 1 つであるが，これらはトピックごとの単語の確率分布である ϕ に影響を受ける．ここでは各著者はトピック分布で特徴つけられると仮定しており，テスト著者のトピック分布と類似しているものを探せばよい．ある著者のトピックの確率分布 θ とある著者が正解であるテスト文書のトピックの確率分布 θ が類似し，トピックの確率分布は著者の特徴を表すものとなり，テスト文書のトピックの確率分布 θ は正解である著者のトピック確率分布 θ と最も類似すると考えられる．よって，本研究では，2 つのトピックの確率分布を比較することにより著者推定を行う．トピックの確率分布を比較する方法として，2 つの確率分布の差の尺度である KL 情報量と 2 つの分布がどのくらい独立しているかを調べるカイ 2 乗値を用いる．θ_x を著者 x のトピックの確率分布，θ_d をテスト文書 d のトピックの確率分布として，トピックの確率分布の KL 情報量を次のように定義する．

$$KL(\theta_x // \theta_d) = \sum_i \theta_{xi} log \frac{\theta_{xi}}{\theta_{di}}$$

また，トピックの確率分布の X^2 値を次のように定義する．

$$X^2 = \sum_i \frac{(\theta_{di} - \theta_{xi})^2}{\theta_{xi}}$$

2.4 実験

2.4.1 実験準備

実験には 2 つのコーパスを用いる．1 つ目のコーパスには，夏目漱石，森鴎外，島崎藤村の 3 人の著者の小説を 10 作品ずつ計 30 作品 (表 3.1) を用いる． 各小説は，1 章と 2 章を学習データとし，3 章以降をテストデータとする．2 つ目のコーパスには，2007 年度の毎日新聞，朝日新聞，読売新聞の社説を用いる．各社説は，9ヵ月分を学習データ，3ヵ月分をテストデータとする．社説の総数を表 2.2 に示す．実験データには MeCab による形態素解析を行い，名詞，動詞，形容詞，副詞のみを抽出して用いる．また，各実験データ中で 1 度しか出現しない単語は取り除いて用いる．

これらのコーパスを用いて 4 つの実験を行う．まず最初に第 1 の実験では，パープレキシティを測定することによりモデルの評価を行う．次に第 2 の実験では，小説と社説を用いて著者推定を行う．第 3 の実験では，小説の 3 人の著者の内，夏目漱石，森鴎外のみを学習して 3 人の著者を推定する．そして最後に，第 4 の実験では小説と社説を混合して著者推定を行う．

著者	作品名
夏目 漱石	坊っちゃん, 硝子戸の中, 彼岸過迄, こころ, 草枕, 行人, 道草, 門, 野分, 三四郎
森 鴎外	かのように, 魚玄機, 雁, 最後の一句, 細木香以, 心中, 二人の友, 高瀬舟, 寒山拾得, 安井夫人
島崎 藤村	嵐, 旧主人, 食堂, 岩石の間, 千曲川のスケッチ, 船, 分配, 藁草履, 家 (上), 刺繍

表 2.1: 小説リスト

パラメタ推定には乱数を使用しているため，推定結果は乱数によって結果が異なる．このため，乱数の初期値を変化させて実験を 10 回行い，その平均値を正解率とする．また，ディリクレ分布のパラメタである α，β の値は 0.01 とし，トピック数は小説で 100，社説で 80，小説と社説の混合で 100 とする．ギブスサンプリングの繰り返し回数は 500 回とする．

	毎日	朝日	読売
学習データ	515	507	513
テストデータ	177	171	175

表 2.2: 社説の総数

2.4.2 評価方法

実験の評価方法としてパープレキシティと F 値を用いる．F 値は再現率と精度の調和平均であり，実際に正であるもののうち，正であると予測されたものの割合である再現率を次のように定義する．

$$R_i = \frac{a_i}{a_i + c_i} \tag{2.4}$$

また，正と予測したデータのうち，実際に正であるものの割合である精度を次のように定義する．

$$P_i = \frac{a_i}{a_i + b_i} \tag{2.5}$$

a_i は推定結果が正である数，c_i は正であるが負と推定された数，b_i は正であると推定した中で正解が負である数である．この 2 つの式の調和平均である F 値は次のように定義する

$$F_i = \frac{2 \times P_i \times R_i}{P_i + R_i}, \quad F = Average_i F_i$$

2.4.3 実験結果

まず初めに，小説，社説，小説と社説の混合のテストセットパープレキシティの値を表 2.3 に示す．ここでトピック数は 100，ギブスサンプリングの繰り返し回数は 500 回である．社説のパープレキシティの値は小説の 1.266 倍となっており，もっとも悪い値を示している．

小説	社説	小説＋社説
1140	1443 (1.266)	1393 (1.222)

表 2.3: パープレキシティ（トピック数 100）

著者推定を行った結果の正解率を表 4，表 5 に示す．小説での著者推定の全体の正解率は，KL 情報量で 97%，X^2 値で 92% である．社説では，著者推定の全体の正解率は，KL 情報量で 71%，X^2 値で 59% である．

著者同士のトピックの確率分布の比較結果を表 2.6 と表 2.7 に示す．著者推定の正解率の高くなる小説では著者同士の差が大きくなっていることが確認できる．また，KL

	夏目	森	島崎	全体
(KL)				
再現率	100	94	97	97
精度	92	100	100	97
F 値	96	97	98	97
(X^2)				
再現率	75	100	97	91
精度	100	79	100	93
F 値	85	88	98	92

表 2.4: 正解率（小説）

	毎日	朝日	読売	全体
(KL)				
再現率	57	89	66	71
精度	67	71	75	71
F 値	61	79	69	71
(X^2)				
再現率	65	63	42	57
精度	48	64	71	61
F 値	55	62	51	59

表 2.5: 正解率（社説）

情報量による比較で最も F 値の小さい毎日新聞では他の新聞社との差が小さくなっている．

そして次に，各著者で確率の高いトピックのトップ 10 を表 2.8 と表 2.9 に示す．表 2.8 を見ると，すべての著者に共通して表れているトピックはトピック 84 だけである．表 2.9 では，すべての著者で確率の高くなっているトピックはトピック 76，70，53 となり，共通して確率の高くなるトピックが多い．

第 3 の実験では，夏目漱石と森鴎外のみを学習したことから，夏目漱石でも森鴎外でもなければ島崎藤村と判断する．このため，閾値を設けその値を $KL = 5$，$X^2 = 30000$ とする．著者推定の結果を表 2.10 に示す．正解率は，KL 情報量で 94%，X^2 値で 91% である．

最後の実験では小説の 3 人の著者と社説の 3 つ新聞社の計 6 人の著者で著者推定を行う．その結果を表 2.11 に示す．正解率は，小説では KL 情報量で 95%，X^2 値で 86%，社説では KL 情報量で 67%，X^2 値で 51% である．

ここで 6 人の著者同士の比較結果を表 2.12 に示す. 表 2.12 より小説の著者と社説の

	夏目	森	島崎
(KL)			
夏目	0	5.837	7.558
森	8.045	0	7.486
島崎	8.547	6.296	0
(X^2)			
夏目	0	128372	110690
森	15285	0	32309
島崎	68433	112637	0

表 2.6: 著者同士の比較 (小説)

	毎日	朝日	読売
(KL)			
毎日	0	1.095	1.005
朝日	0.655	0	1.843
読売	0.726	1.410	0
(X^2)			
毎日	0	3276	29486
朝日	26959	0	41355
読売	21002	58470	0

表 2.7: 著者同士の比較 (社説)

著者では，大きく異なっていることが見てとれる．次に各著者で確率の高くなるトピックのトップ 10 を表 2.13 に示す．小説で共通して確率の高くなるトピックはトピック 12 だけであるが，社説ではトピック 37，40，66，75，79 と 5 つのトピックが共通して表れている．

2.4.4 考察

著者推定の正解率は，KL 情報量による比較で小説のときの 97%が最も高い値となり，小説と社説の混合での 67%が最も低い値となっている．また，パープレキシティ値は小さい値となっている．

高い正解率となる小説での著者推定では，表 6，表 7 より各著者に明確な相違ができていることがわかる．また，表 8 より，トピック 84 だけがすべての著者に共通して表れていることがわかる．ここでトピック 84 の内訳を表 2.14 に示す．表 2.14 より，トピック 84 には特徴的な語が出現していない．

夏目		森		島崎	
トピック	確率	トピック	確率	トピック	確率
17	0.1015	34	0.1526	78	0.0931
30	0.0607	28	0.1036	84	0.0905
84	0.0502	45	0.0925	34	0.0864
95	0.0497	76	0.0705	61	0.0741
91	0.0492	7	0.0541	60	0.0608
54	0.0487	84	0.0541	53	0.0592
26	0.0454	17	0.0538	14	0.0588
41	0.0434	1	0.0417	62	0.0498
76	0.0336	55	0.0393	19	0.0353
31	0.0279	96	0.0364	5	0.0330

表 2.8: トピックの確率のトップ 10（小説）

毎日		朝日		読売	
トピック	確率	トピック	確率	トピック	確率
76	0.0692	29	0.0727	76	0.0724
75	0.0522	27	0.0643	54	0.0594
13	0.0465	13	0.0483	75	0.0520
40	0.0437	70	0.0427	26	0.0451
27	0.0428	53	0.0410	70	0.0382
70	0.0420	36	0.0384	72	0.0375
64	0.0415	76	0.0363	5	0.0340
72	0.0352	46	0.0342	40	0.0338
10	0.0334	47	0.0331	7	0.0331
53	0.0290	14	0.0306	53	0.0321

表 2.9: トピックの確率のトップ 10（社説）

一方，社説では著者同士の相違が少なくなっている．特に毎日新聞と朝日新聞では，KL 情報量の値が非常に小さい値となっており，毎日新聞から朝日新聞を区別することは困難となっている．これは，著者に共通して表れるトピックが多くなっているためにF 値が減少していると考えられる．しかしながら，社説であっても著者の個性によって著者推定が行えている．

KL 情報量は情報の量であるのに対し，X^2 値は誤差の量である．このため，得られるF 値は異なるが，小説や社説の著者推定では最もF 値が高くなる著者は同じになっている．表 6 より小説の著者間には明確な特徴が表れていることから，著者トピックモデルにより著者に明確な特徴が与えられ，著者推定が行えていると考えられる．

小説と社説の混合のテストセットパープレキシティは社説の値より小さい値となっており，正解率は KL 情報量による比較で小説が 95%，社説が 67%である．また、表 2.12 より，小説と社説には明確な相違があることがわかる．表 2.13 では小説と社説で共通して表れるトピックがなく，2 つの領域を正確に分離できていると考えられる．

これらの結果から，領域に依存するトピックを検出し，これらが共通するものではないことから，著者推定が領域に独立して動作すると考えられる．

	夏目	森	島崎	全体
(KL)				
再現率	97	91	92	93
精度	88	100	95	94
F 値	92	95	93	94
(X^2)				
再現率	74	100	96	90
精度	100	82	94	92
F 値	85	90	95	91

表 2.10: 正解率（小説）

	夏目	森	島崎	小説	毎日	朝日	読売	社説
(KL)								
再現率	100	93	92	95	51	81	69	67
精度	87	100	100	96	62	74	65	67
F 値	93	96	96	95	56	77	66	67
(X^2)								
再現率	88	100	92	93	41	61	50	50
精度	92	48	100	80	48	55	54	53
F 値	89	63	96	86	43	57	51	51

表 2.11: 正解率（混合）

2.5 結論

本論文では，LDA の構造内で著者推定の領域独立を示した．すなわち，トピックが著者間の潜在的な区別できる特性を捉えることから，著者トピックモデルは領域カテゴリーから独立していることを経験的に示した．実験を通して，著者トピックモデルは (1) 各著者に対して明確な単語の分布を表すいくつかのトピックを検出する，(2) 一般的でない文書でもトピックは対象のカテゴリーに依存する，(3) これらのトピックの混合の意味によって動作する，ということが言える．

	夏目	森	島崎	毎日	朝日	読売
(KL)						
夏目	0	3.256	3.927	13.018	11.373	13.606
森	5.274	0	4.653	13.442	12.591	13.688
島崎	5.879	4.838	0	14.096	12.996	13.019
毎日	10.133	10.471	11.374	0	0.620	0.954
朝日	10.615	10.209	11.282	0.467	0	3.173
読売	9.804	10.599	11.547	0.338	0.654	0
(X^2)						
夏目	0	73341	56024	86067	101628	75854
森	6774	0	23866	54325	46076	67472
島崎	17809	41944	0	137839	118245	171731
毎日	851838	1231037	1886942	0	141	6
朝日	834555	1252465	1636291	16485	0	18
読売	1081328	1279498	860258	27994	322294	0

表 2.12: 著者同士の比較（混合）

夏目		森		島崎		毎日		朝日		読売	
トピック	確率	トピック	確率	トピック	確率	トピック	確率	トピック	確率	トピック	確率
32	0.1395	62	0.1300	87	0.2096	37	0.1398	43	0.0936	37	0.1873
24	0.0982	32	0.1158	9	0.1117	36	0.0958	92	0.0918	46	0.0827
34	0.0792	70	0.1096	52	0.0972	17	0.0705	37	0.0849	40	0.0765
9	0.0668	49	0.0925	67	0.0792	40	0.0675	75	0.0576	36	0.0652
4	0.0569	12	0.0711	98	0.0637	75	0.0559	40	0.0522	26	0.0584
91	0.0523	65	0.0625	12	0.0543	79	0.0481	66	0.0515	99	0.0450
96	0.0521	4	0.0543	27	0.0515	26	0.0416	79	0.0514	66	0.0429
31	0.0514	45	0.0529	100	0.0454	46	0.0412	17	0.0483	79	0.0402
100	0.0426	24	0.0464	25	0.0442	66	0.0401	36	0.0458	75	0.0331
12	0.0402	31	0.0306	71	0.0407	97	0.0297	15	0.0329	84	0.0287

表 2.13: トピックの確率のトップ 10（混合）

単語	確率
よう	0.1126
いる	0.1085
時	0.0398
もの	0.0371
ない	0.0361
そう	0.0338
思う	0.0332
学校	0.0227
くれる	0.0211
一	0.0179

表 2.14: トピック 84（小説）

第3章　確率モデルによる品詞分布の特徴推定

本研究では名詞の割合の事前分布にガウス分布を用いたジャンル分類を提案する．単語の頻度を用いる分類では数万次元の単語分布を比較する必要があったのに対し，本手法では品詞分布により分類を行うため高速に分類が行える．日本語文書の名詞に対する他の品詞の割合には相関関係があることが知られており，いくつかの近似式が示されている．品詞分布の特性は文書のジャンルごとに表れると仮定し，ジャンル分類を行う．

3.1　前書き

計量言語学は，言語現象を統計手法を用いて定量的に分析する分野である．書き言葉である文章テキストは言語現象の例であり，計量文体学や計量的コーパス言語学などは計量言語学として盛んに論じられている．

文章は，何らかの文字列が一定の文法規則に基づいた文の集合体であり，定型化されていないデータとして典型的である．記号列が何らかの規則に従ってはいるが，定型化されていないため，単語やフレーズ，品詞など何らかの単位に分割し，それらの出現頻度や共起関係・同時出現などを抽出して，定量的に解析できる．この手法は総称してテキストマイニングと呼ばれている．計量文体分析の分野では一世紀以上も前から，文章を構成する要素の特徴を定量的に分析する方法で文章の執筆者の推定などを行っている．例えば，一般人が書いた短い文章の書き手の同定・推定では短い文章でも書き手がほぼ同定できる [14]．このような手法は，手紙や脅迫文の著者同定，スパムメール識別，ブログ分類にも応用されている．もっと一般的は応用分野として，アンケート分析などを用いた顧客サービスの自動評価，テキストマネジメントと情報・知識の抽出などがある．

文書は，その目的や対象とする読者のためにある統一的な難易度で作成され，例えば“ 増える ”という日常表現は，論文などの公式的な分野では“ 増加する ”という表現に代わって利用される．このように，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．

本研究では延べ語数での品詞分布を用いることで文書のジャンル類別が行えること

を示す．品詞分布を特徴量として用いてジャンル分類が高精度に実施可能ならば，次元数が品詞の種類数となり，小さい次元で文書の比較を行うことができる．この結果，次元数が極めて少ないことでデータ数が増えても必要なメモリ量は少量で済むという利点がある．

第 2 章ではジャンル分類について述べ，第 3 章は日本語文書の品詞分布特性を論じる．第 4 章では提案手法について述べ，第 5 章では実験によりこの有効性を示す．第 6 章は結論である

3.2 日本語文書のジャンル分類

文書を構成する文章テキストには，その文書ジャンルに含まれる典型的な構成要素，テキストにおける構成要素の一般的な出現順序，構成要素間の関係などに特徴があるとされる．ジャンルに応じた特徴を利用し，文書の属するジャンルを推定することにより，当該文書の重要箇所抽出や要約文生成，表現の書き換えなどに重要な手掛かりを与えることになる．たとえば，論文ならば目的，方法，結果，考察，結論などの構造が，手紙なら，頭語，季節の挨拶，安否の問いかけ，本文，結語などから構成される．また，”文体” (style) とは，それぞれの作者がジャンルに応じて特有な文章表現上の特色を有し，文章の語句・語法・修辞などが作者固有の特徴・傾向となっているものをいう．

本研究では延べ語数での品詞分布を用いることで文書のジャンル類別が行えることを示す．文書の特徴としては語の分布が挙げられ，ジャンル分類では通常語の頻度が用いられる．小説や新聞などの系列長の長い文書では語彙数が数万から数十万となり，語の頻度による分類では数万次元の単語分布の比較が必要となる．これによりデータ数の増加に伴い必要なメモリ量が膨大となる．また，単語を特徴量として用いると語の共起による分類では多義語などの語の意味の扱いや，語の確率による分類では出現しない語の扱いにより分類結果が異なるという問題がある．

しかし，特徴量として品詞分布を用いることで，ジャンル分類に必要な次元数が語彙数から品詞数へと大幅に削減でき語の扱いが容易になる．また，日本語文書の品詞分布には法則性があることが知られており [11][12]，各ジャンルの品詞の割合の理論値は品詞分布の決定要素である名詞の割合によって求めることが可能である．文章を構成する主語・述語には，修飾語などの要素が付け加わって，より複雑な構造を形成する．文章を成り立たせる要素を「文の成分」というが，通常「主語」「述語」「修飾語」(連用修飾語・連体修飾語)「接続語」「独立語」の５つが挙げられる．形態素的には，名詞や動詞など独立した意味を担う自立語と，格などでこれらを補完する補助語があり，個々の語の利用を調べるだけではなくて，これらの分布を特徴量として用いることで，計量的な判断を高精度にすることが考えられる．

この法則性を用いることで品詞分布により各ジャンルの特徴を表すことができ，ジャンル分類が行えると考えられる．ジャンル分類では，各ジャンルに確率モデルを与え，

品詞分布の生成過程を表現する．名詞の割合の事前分布にガウス分布を仮定し，ジャンルごとにガウス分布のパラメータを学習することで各ジャンルの名詞分布の特徴を得る．このガウス事前分布による各ジャンルへの所属確率と確率モデルから求まる理論値と観測値の誤差によりジャンル分類が行えることを示す.

3.3 日本語文書の品詞分布

日本語文書では，自立語と付属語が組み合わされて出現する．自立語とは独立して意味を成す語であり，典型的には名詞や動詞がある．一方付属語とは単独では意味を持たず自立語と合わせて節を成す語であり，助詞や助動詞が挙げられる．例えば「出た出た月が」を単位語に区切ると「出/た/出/た/月/が」となり，自立語は「出」と「月」，延べ語数は 6 である.「出る」が 2 ,「た」が 2 ,「月」が 1 ,「が」が 1 であり，これはそれぞれの語の使用度数を示す．単位語の総数から，重複するものを省いて数えた語のが異なり語である．この例では，延べ語数は 6 であるが,「出る」と「た」が重なっているため，異なり語数は「出る」「た」「月」「が」の 4 つである.

樺島の法則は，延べ語数に関する品詞分布の特性である．自立語を名詞，動詞，名詞・動詞などについてそれがどのようであるかというありさまを示す形容詞類（形容詞・副詞・連体詞），文の成分としては比較的独立性をもつ接続詞類（接続詞・感動詞）に分ける．このとき，名詞の百分率を N，動詞を V，形容詞類を A_d，接続詞類を I としたとき各品詞の割合は以下の近似式で表される.

$$A_d = 45.67 - 0.60N \tag{3.1}$$

$$logI = 11.57 - 6.56logN \tag{3.2}$$

$$V = 100 - (N + A_d + I) \tag{3.3}$$

樺島の法則で示されている近似式は名詞の割合に対しての品詞分布の特性である．品詞の割合には文書のジャンルによって差がある [10] としながら，ジャンルによる品詞分布の特性を考慮していないため，当てはまりがあまり良くない.

文書ジャンルを考慮した品詞分布の法則としては，古典文学作品での品詞分布に関する大野の語彙法則がある．大野の語彙法則は異語数に関するものではあるが，古典 9 作品において名詞の割合がもっとも高い文書を左端に，名詞の割合がもっとも低い文書を右端におき二つの作品の品詞の割合を直線で結ぶと，他の作品の品詞の割合はこの直線状に垂直に並ぶ．ただ，大野の研究では古典文学作品のみに焦点を当てており，他のジャンルについてこの法則が成り立つかは論じられていない.

大野は語彙法則をグラフで図示したが，この定式化が水谷によってなされた [13]．ある 3 作品の名詞構成比を X_0，x，X_1 他の語類の構成比を Y_0，y，Y_1 としたとき以下の式が成り立つ.

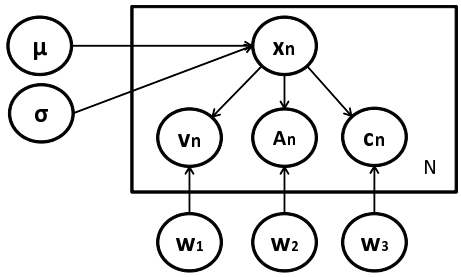

図 3.1: グラフィカルモデル

$$y = Y_0 + \frac{Y_1 - Y_0}{X_1 - X_0}(x - X_0) \tag{3.4}$$

先行研究が着目するポイントは，延べ語数，異語数に関して日本語文書には品詞分布に法則性が存在することにある．本研究では品詞分布の特性がジャンルごとに変化すると仮定し，文書のジャンルごとに回帰直線を設ける．これにより品詞分布での分類が名詞により相関分析でき，名詞分布によりジャンル分類が行える．

3.4 提案手法

3.4.1 品詞分布の確率モデル

本稿では，品詞分布の対応関係を確率モデルで表現し，ジャンル分類に用いる (図 3.1 で表す)．ここで X は名詞の割合，V は動詞の割合，A は形容詞類の割合，C は接続詞類の割合である．w は多項式係数ベクトルであり，μ と σ はガウス事前分布のパラメータである．文書の名詞の割合は事前分布であるガウス分布によって決まり，ガウス事前分布のパラメータがジャンルごとの名詞分布の特徴を表す．動詞，形容詞類，接続詞類の割合は名詞の割合と多項式係数ベクトルによって求まる．この確率モデルを各ジャンルに与え，ジャンルごとにガウス分布のパラメータと多項式係数ベクトルの学習を行う．ガウス分布のパラメータは学習データの平均と分散より (5) 式で求まる．各品詞の多項式係数ベクトル w_i は $\{w_{i0}, w_{i1}\}$ で表され，(6) 式・(7) 式より求まる．

$$\mu = \frac{1}{n}\sum_{n=1}^{N} x_n \quad , \quad \sigma = \frac{1}{n}\sum_{n=1}^{N}(x_n - \mu)^2 \tag{3.5}$$

$$w_{i0} = \frac{n\sum_{k=1}^{n} x_k y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k}{n\sum_{k=1}^{n} x_k^2 - (\sum_{k=1}^{n} x_k)^2} \tag{3.6}$$

$$w_{i1} = \frac{\sum_{k=1}^{n} x_k^2 \sum_{k=1}^{n} y_k - \sum_{k=1}^{n} x_k y_k \sum_{k=1}^{n} x_k}{n \sum_{k=1}^{n} x_k^2 - (\sum_{k=1}^{n} x_k)^2} \tag{3.7}$$

3.4.2 ジャンルの分類

ジャンル分類では，ジャンルごとに学習したパラメータより，動詞，形容詞類，接続詞類の品詞の割合の観測値と理論値の誤差を求め，ガウス分布を仮定した名詞の出現確率の逆数をかけることによりジャンル分類を行う．これにより単語の頻度を用いる分類では数万次元の単語分布を比較する必要があったのに対し，本手法では各ジャンルの名詞の出現確率と多項式係数ベクトルにより分類を行うため高速に分類が行える．各ジャンルでの名詞の出現確率はガウス事前分布のパラメータである μ_j，σ_j により以下の式で求まる．

$$N(j) = \frac{1}{(2\pi\sigma_j)^{\frac{1}{2}}} exp(-\frac{1}{2\sigma_j}(x - \mu_j)^2) \tag{3.8}$$

各ジャンルの動詞，形容詞類，接続詞類の理論値はテスト文書 $\{d_1, d_2, ..., d_n\}$ の名詞の割合 $\{x_1, x_2, ..., x_n\}$ と多項式係数ベクトルによって $w_{ji0} + w_{ji1} * x_n$ で求まる．ここで j はジャンル数であり，i は品詞数である．テスト文書の品詞の割合を動詞 $\{v_1, v_2, ..., v_n\}$, 形容詞類 $\{a_1, a_2, ..., a_n\}$, 接続詞類 $\{c_1, c_2, ..., c_n\}$ としたとき，理論値との誤差は以下の式より求まる．

$$\begin{aligned} V(j) = &(v_n - (w_{j11}x_n + w_{j10}))^2 + (a_n - (w_{j21}x_n + w_{j20}))^2 \\ &+ (c_n - (w_{j31}x_n + w_{j30}))^2 \end{aligned} \tag{3.9}$$

ジャンル分類では品詞分布の理論値と観測値の誤差と名詞の出現確率の逆数によって求まる値が最も低いジャンルを分類結果とする．以下の式に基づいて分類を行う．

$$Ans = arg \min_j V(j) * \frac{1}{N(j)} \tag{3.10}$$

3.5 実験

本章では2つの実験を行う．第1の実験ではF検定を行うことにより，回帰式が品詞分布の特性を表すものであることを示す．回帰式がジャンルに固有のものとなればコーパスに依存せずジャンル分類が行える．第2の実験ではジャンル分類を行う．ジャンル分類ではベースライン手法として単純ベイズを用いて精度と実行時間の比較を行う．

品詞分布による分類が分類法によらず行えるか検証する．また，(形態素ではなくて) 語自体の頻度による分類として，名詞のみを対象とする Latent Direchlet Allocation (LDA) を用いる．LDA による分類とは，ジャンルのトピックの確率分布とテスト文書のトピックの確率分布を比較し，KL 情報量が最も小さいジャンルを推定結果とするものである [15]．単純ベイズでは尤度が最も高いジャンルを推定結果とする．

3.5.1 実験準備

実験には 4 つのコーパスを用いる．各コーパスは，5 人の著者の小説を 20 作品ずつ計 100 作品 (表 3.1)，朝日新聞の 2007 年 1 月度を 30 日分，日本語話し言葉コーパス (Disc3) を収録順に先頭から 100 文書，NTCIR-3 より 98 年度公開特許公報全文データを 100 文書 (表 3.2) である．特許データは発明の詳細な説明のみを抽出して用いる．また，これらのうち小説・話し言葉・特許は 50 文書，新聞は 20 日分を学習データとして用いる．実験データには茶筅による形態素解析を行う．比較手法である LDA のパラメータはトピック数 150，ディリクレ分布の初期値を α=0.01，β=0.01 とし，ギブスサンプリング繰返しは 500 回行う．実行時間の測定に用いる計算機は CPU Intel Corei3 1.33GHz，メモリ 4GB である．

著者	作品名
夏目 漱石	坊っちゃん，文芸と道徳，ケーベル先生，硝子戸の中 彼岸過迄，一夜，こころ，草枕，行人，幻影の盾 道草，門，手紙，野分，三四郎，変な音 趣味の遺伝，二百十日，虚子君へ，落第
芥川 龍之介	アグニの神，或敵打の話，桃太郎，疑惑，犬と笛 歯車，英雄の器，二つの手紙，一夕話，報恩記 影，羅生門，開化の良人，河童，三つの窓 おしの，地獄変，蜃気楼，運，彼
太宰 治	老ハイデルベルヒ，玩具，グッド・バイ，走れメロス 逆行，女生徒，佳日，火の鳥，鴎，犯人 喝采，故郷，狂言の神，黄金風景，ろまん燈籠 正義と微笑，斜陽，嘘，兄たち，女の決闘
島崎 藤村	秋草，嵐，朝飯，旧主人，食堂，岩石の間 三人の訪問者，並木，千曲川のスケッチ，船 伸び支度，芽生，分配，ある女の生涯，藁草履 家 (上)，刺繍，再婚について，北村透谷の短き一生，家（下）
森 鴎外	あそび，興津弥五右衛門の遺書，阿部一族，かのように，雁 カズイスチカ，魚玄機，最後の一句，細木香以，じいさんばあさん 堺事件，文づかい，余興，みちの記，心中 食堂，二人の友，沈黙の塔，高瀬舟，鶏

表 3.1: 小説リスト

公開番号
特開平１０－１，...，特開平１０－１００

表 3.2: 特許リスト

3.5.2 評価方法

回帰直線の評価にはF検定を用い，分類の評価にはf尺度を用いる．F検定では，回帰による要因効果と誤差による変動要因から求まる値が有意水準以上であれば，回帰直線と観測値の間に有意差が有り，回帰直線が品詞分布の特性を表しているといえる．ここで回帰による要因効果についての平方和 S_F と平均平方 V_F を以下の式で定義する．

$$S_F = \sum_{i=1}^{n} (\bar{Y} - Y_i)^2 \quad , \quad V_F = \frac{S_F}{m-1} \tag{3.11}$$

次に誤差による変動要因についての平方和 S_C と平方平均 V_C を以下の式で定義する．

$$S_C = \sum_{i=1}^{n} (y_i - Y_i)^2 \quad , \quad V_C = \frac{S_C}{n-m} \tag{3.12}$$

この時,

$$F = \frac{V_F}{V_C} \tag{3.13}$$

となる．自由度は分母に対して (全標本数-群数) であり，分子に対して (群数-1) である．

次に分類の評価方法としてf尺度を用いる．f尺度は再現率と適合率の調和平均であり，実際に正であるもののうち，正であると予測されたものの割合である再現率を次のように定義する．

$$R_i = \frac{a_i}{a_i + c_i} \tag{3.14}$$

また，正と予測したデータのうち，実際に正であるものの割合である適合率を次のように定義する．

$$P_i = \frac{a_i}{a_i + b_i} \tag{3.15}$$

a_i は推定結果が正である数，c_i は正であるが負と推定された数，b_i は正であると推定した中で正解が負である数である．この2つの式の調和平均であるf尺度を次のように定義する

$$f_i = \frac{2 \times P_i \times R_i}{P_i + R_i}, \quad f = Average_i f_i$$

	小説	話し言葉	特許	新聞
動詞				
w_{11}	-0.461	-0.039	-0.815	-0.690
w_{10}	0.557	0.304	0.804	0.718
形容詞類				
w_{21}	-0.453	-0.368	-0.127	-0.276
w_{20}	0.375	0.297	0.132	0.251
接続詞類				
w_{31}	-0.086	-0.593	-0.057	-0.033
w_{30}	0.069	0.399	0.064	0.031

表 3.3: 回帰式の係数

3.5.3 実験結果

まず，回帰式の係数とガウス分布のパラメータを表 3.3，3.4 に示す．各パラメータはジャンルごとに異なっており，ジャンルの特徴を表している．F 検定の結果を表 3.5，3.6 に示す．すべての値が有意水準 5 %での F 分布の基準値より大きい値となっていることから，有意水準 5 %で回帰直線と観測値には有意差が有り，回帰直線が品詞分布の特性を表していると言える．

	小説	話し言葉	特許	新聞
平均	0.585	0.440	0.779	0.799
分散	1.84E-03	5.07E-03	1.63E-03	1.12E-04

表 3.4: 各ジャンルの名詞分布の平均と分散

	動詞	形容詞類	接続詞類
小説	61.94	19.77	58.08
話し言葉	33.27	54.08	88.11
特許	364.68	85.50	45.49

表 3.5: F 検定結果 (自由度 1,48)

次にジャンル分類の結果と実行時間を表 3.7，3.8 に示す．ジャンル分類の f 尺度は提案手法で 0.945，単純ベイズで 0.899，LDA で 0.99 となる．(単語分布を用いた) LDA による分類と比して，品詞分類を用いた手法では精度が低いが, 提案手法では遜色がないほど高精度に分類が行える．提案手法による分類ではすべてのジャンルの f 尺度が高

	動詞	形容詞類	接続詞類
新聞	85.40	8.38	12.96

表 3.6: F 検定結果 (自由度 1,8)

い値を示す．一方，単純ベイズでは新聞の f 尺度が低い．LDA 分類と比較すると，単語分布では新聞と特許の f 尺度が比較的低いが，品詞分布による分類では逆に新聞と特許の f 尺度が他のジャンルより高いことに注目したい．実行時間は，提案手法で 5.05 秒，単純ベイズで 4.85 秒，LDA で 24576 秒となる．提案手法と単純ベイズではほとんど差がないが，LDA では品詞分布による分類の約 5000 倍となっている．

	新聞	小説	話し言葉	特許	全体
(提案手法)					
再現率	1	1	0.840	0.960	0.950
適合率	0.909	0.847	1	1	0.939
f 値	0.952	0.917	0.913	0.980	0.945
(単純ベイズ)					
再現率	1	1	0.82	0.88	0.925
適合率	0.667	0.833	1	1	0.875
f 値	0.800	0.909	0.901	0.936	0.899
(LDA)					
再現率	0.960	1	1	1	0.990
適合率	1	1	1	0.962	0.990
f 値	0.980	1	1	0.980	0.990

表 3.7: ジャンル分類結果

	提案手法	単純ベイズ	LDA
学習時間	4.92	4.74	24575
分類時間	0.125	0.109	0.68
合計	5.05	4.85	24576

表 3.8: 実行時間 (秒)

	小説	話し言葉	特許	新聞
群内分散	4.04E-04	1.25E-03	1.91E-04	8.09E-06

表 3.9: 動詞の群内分散

3.5.4 考察

ジャンル分類では提案手法による分類で f 値が 0.945 と高い精度になっている．単純ベイズによる分類では特許との名詞の割合の平均値が近い新聞で f 値が小さくなっているが，提案手法ではすべてのジャンルで高い精度となっている．このため，提案手法では名詞の割合の分布が近いジャンルが存在してもジャンル分類が行えると考えられる．また，単語分布による分類においても新聞と特許の f 尺度は他のジャンルより低くなったが，提案手法では逆に他のジャンルより高くなっている．表 3.9 よりテスト文書の動詞の群内分散を見ると，F 検定の値が大きい特許と新聞では分散が 1.91E-04, 8.09E-06 となり非常に小さい値となっている．分散が最も大きい話し言葉では分散が 1.25E-03 となり，この値は分散の最も小さい新聞の 155 倍である．提案手法で f 尺度の高くなったジャンルは群内分散の小さいジャンルであり，回帰式の当てはまりがいいために精度が高くなったと考えられる．分類の実行時間は品詞分布での分類に比べて，繰り返し学習が必要な LDA での分類は 5000 倍近い値となっており，品詞分布での分類は非常に高速に行える．

3.6 結論

本論文では，品詞分布の決定要素である名詞分布によるジャンル分類法を提案した．また，品詞分布の特性は文書のジャンルごとに異なっており，ジャンルごとの回帰直線を用いることにより品詞分布の特性が得られることを検定により示した．

第4章　品詞分布を用いた日本語文書のジャンル分類

本研究では名詞の割合の事前分布にガウス分布を用いたジャンル分類を提案する．単語の頻度を用いる分類では数万次元の単語分布を比較する必要があったのに対し，本手法では品詞分布を用いるため品詞の種類数の次元で分類が行える．これにより，ストリームデータのように各ジャンルの特徴が逐次変化する場合でもパラメータの更新が容易となる．日本語文書の名詞に対する他の品詞の割合には相関関係があるため、各ジャンルの特徴として用いる．実験により有効性を示す.

4.1　前書き

近年，インターネットの発達から大量の文書を入手することが可能になり，静的な文書集合に対する分類だけでなく，次々に新たな文書が到着するという文書ストリームに対する分類手法の重要性が高まっている．ストリームデータは新しいデータが逐次到着することからデータ量が膨大となる．このことから文書ストリームの分類には限られた特徴量で高速に分類する手法が必要となる．ジャンル分類では通常語の頻度が用いられ，ストリームデータの分類に対しても語の頻度を特徴量とした分類法が提案されている [16][17]．語の頻度を用いた場合，系列長の長い文書では語彙数が数万から数十万となり，数万次元の単語分布の比較が必要となる．また，データ数が増加していくことから必要なメモリ量が膨大となる．さらに，ストリームデータでは文書集合が動的に変化するため，ジャンルの特徴となる語が変わり，特徴語を選択することが困難になる．しかし，特徴量として品詞分布を用いることで，ジャンル分類に必要な次元数が語彙数から品詞数へと大幅に削減でき語の扱いが容易になる．これによりストリームデータのように各ジャンルの特徴が逐次変化する場合でもパラメータの更新が容易に行える．単語分布は変化しても用いる品詞は同じであることから，品詞分布の変化としてジャンルの特徴の変化を捉えることができる．本研究では，品詞分布を用いて様々なジャンルの文書が次々に到着すると仮定した文書ストリームに対する分類法を提案する．日本語文書の品詞分布には法則性があることが知られており [11]，各ジャンルの品詞の割合の理論値は品詞分布の決定要素である名詞の割合によって求めることが可能である．この品詞分布の法則性と各ジャンルの名詞の割合の事前分布にガウス分布を用いることでジャンル分類が行えることが示されている [18]．各ジャ

ンルのガウス分布のパラメータと多項式係数ベクトルを更新することにより文書ストリームの分類を行う．

第 2 章では提案手法について述べ，第 3 章では実験により有効性を示し，第 4 章で結論とする．

4.2 提案手法

4.2.1 分類方法

ジャンル分類では各ジャンルの名詞の割合の事前分布であるガウス分布と，品詞分布の線形近似の係数である多項式係数ベクトルを用いて分類を行う．文書の名詞の割合は事前分布であるガウス分布によって決まり，ガウス事前分布のパラメータがジャンルごとの名詞分布の特徴を表す．動詞，形容詞類，接続詞類の割合は名詞の割合と多項式係数ベクトルによって求まる．各品詞の多項式係数ベクトル (w_{i0}, w_{i1}) は以下の式より求まる．

$$w_{i0} = \frac{n\sum_{k=1}^{n} x_k y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k}{n\sum_{k=1}^{n} x_k^2 - (\sum_{k=1}^{n} x_k)^2} \tag{4.1}$$

$$w_{i1} = \frac{\sum_{k=1}^{n} x_k^2 \sum_{k=1}^{n} y_k - \sum_{k=1}^{n} x_k y_k \sum_{k=1}^{n} x_k}{n\sum_{k=1}^{n} x_k^2 - (\sum_{k=1}^{n} x_k)^2} \tag{4.2}$$

ジャンルごとに学習したパラメータより，動詞，形容詞類，接続詞類の品詞の割合の観測値と理論値の誤差を求め，ガウス分布を仮定した名詞の出現確率の逆数をかけることによりジャンル分類を行う．各ジャンルの名詞の出現確率はガウス事前分布のパラメータである μ_j と σ_j により以下の式で求まる．

$$N(j) = \frac{1}{(2\pi\sigma_j)^{\frac{1}{2}}} exp(-\frac{1}{2\sigma_j}(x-\mu_j)^2) \tag{4.3}$$

各ジャンルの動詞，形容詞類，接続詞類の理論値はテスト文書 $\{d_1, d_2, ..., d_n\}$ の名詞の割合 $\{x_1, x_2, ..., x_n\}$ と多項式係数ベクトルによって $w_{ji0} + w_{ji1} * x_n$ で求まる．ここで j はジャンル数であり，i は品詞数である．テスト文書の品詞の割合を動詞 $\{v_1, v_2, ..., v_n\}$, 形容詞類 $\{a_1, a_2, ..., a_n\}$, 接続詞類 $\{c_1, c_2, ..., c_n\}$ としたとき，理論値との誤差は以下の式より求まる．

$$\begin{aligned} V(j) = &(v_n - (w_{j11}x_n + w_{j10}))^2 + (a_n - (w_{j21}x_n + w_{j20}))^2 \\ &+ (c_n - (w_{j31}x_n + w_{j30}))^2 \end{aligned} \tag{4.4}$$

品詞分布の理論値と観測値の誤差と名詞の出現確率の逆数によって求まる値が最も低いジャンルを分類結果とする．以下の式に基づいて分類を行う．

$$Ans = arg \ \min_{j} V(j) * \frac{1}{N(j)} \tag{4.5}$$

4.2.2 文書ストリームの分類

文書ストリームの分類を行うため，ジャンルの特徴の変化に応じてガウス分布のパラメータと多項式ベクトルを更新する．パラメータの更新は各ジャンルに用意したサンプル文書 T$t_1, t_2, \cdots, t_n$ を用いて，F 検定による回帰直線の当てはまりを求め，更新後の F 値が高くなる場合にパラメータの更新を行う．サンプル文書にはエラー保証を行い，各品詞の割合は理論値から誤差 ε 以内とする．理論値を K としたき，サンプル文書の割合は

$$K - \varepsilon \geq k_t \geq K + \varepsilon \tag{4.6}$$

となる．これによりジャンルの特徴と異なる文書がサンプル文書となることを防ぐ．提案手法ではテスト文書の品詞分布が得られれば分類が行えるため，到着した文書を一定の長さの区間で区切り，先頭の区間のみを分類に用いる．品詞分布を求める際，低頻度の単語がノイズとなり，品詞の割合の誤差が大きくなることから頻度 1 の単語を除いて品詞の割合を求める．この品詞の割合より，分類手法を用いて所属するジャンルを求める．次にサンプル文書を用いて F 検定を行い，更新前のパラメータと更新後のパラメータの比較を行う．F 検定では回帰による要因効果と誤差による変動要因により回帰直線の当てはまりの良さを求める．ここで回帰による要因効果についての平方和 S_F，平均平方 V_F と誤差による変動要因についての平方和 S_C，平方平均 V_C 以下の式で定義する．

$$S_F = \sum_{i=1}^{n} (\bar{Y} - Y_i)^2 \quad , \quad V_F = \frac{S_F}{m-1} \tag{4.7}$$

$$S_C = \sum_{i=1}^{n} (y_i - Y_i)^2 \quad , \quad V_C = \frac{S_C}{n-m} \tag{4.8}$$

この時,

$$F = \frac{V_F}{V_C} \tag{4.9}$$

となる．ここで自由度は分母に対して (全標本数－群数) であり，分子に対して (群数－ 1) である．更新後の F 値が高く，サンプル文書のうち一定数以上が新しい理論値から ε 以内に収まっていればパラメータの更新を行う．また，更新されたとき理論値から ε 以内に収まっていないサンプル文書は除去する．分類されたテスト文書の品詞

の割合が誤差以内に収まっていればサンプル文書に加え，最も古いサンプル文書は取り除く．

4.3 実験

本章では，文書ストリームの分類を行い提案手法の有効性を示す．比較手法としてパラメータを更新しない手法を用いる．

4.3.1 実験準備

実験には 4 つのコーパスを用いる．各コーパスは，5 人の著者の小説を 20 作品ずつ計 100 作品，朝日新聞の 2007 年度を 1 月 1 日から 100 日分，日本語話し言葉コーパス (Disc3) を収録順に先頭から 100 文書，NTCIR-3 より 98 年度公開特許公報全文データを 100 文書である．特許データは発明の詳細な説明のみを抽出して用いる．また，これらのうち各 50 文書を学習データとして用いる．実験データには茶筅による形態素解析を行う．　提案手法のパラメータの値は，区間の長さ 2000 単語，最大サンプル文書数 10，更新時の最低サンプル文書数 5，理論値からの誤差 ε =0.1 とする．

4.3.2 評価方法

実験の評価には f 尺度を用いる．f 尺度は再現率と適合率の調和平均であり，実際に正であるもののうち，正であると予測されたものの割合である再現率と，正と予測したデータのうち，実際に正であるものの割合である適合率を次のように定義する．

$$R_i = \frac{a_i}{a_i + c_i} \tag{4.10}$$

$$P_i = \frac{a_i}{a_i + b_i} \tag{4.11}$$

a_i は推定結果が正である数，c_i は正であるが負と推定された数，b_i は正であると推定した中で正解が負である数である．この 2 つの式の調和平均である f 尺度を次のように定義する

$$f_i = \frac{2 \times P_i \times R_i}{P_i + R_i}, \quad f = Average_i f_i$$

4.3.3 実験結果

実験結果を表 4.1 に示す．表 4.1 より提案手法の f 値は 0.93 であり，比較手法の f 値は 0.898 である

	提案手法	比較手法
再現率	0.93	0.895
適合率	0.936	0.901
f 値	0.933	0.898

表 4.1: 分類結果

4.3.4 考察

表 4.1 より，提案手法と比較手法の f 値を比較すると提案手法の f 値が高くなっている．また，表 4.2，表 4.3 より各ジャンルの多項式係数ベクトルが更新されていることが確認できる．これらの結果から，パラメータを更新することで精度が向上していると考えられる．

	小説	新聞	特許	話し言葉
w_{v0}	-0.57	-0.785	-0.784	0.05
w_{v1}	0.619	0.797	0.782	0.264
w_{a0}	-0.328	-0.2	-0.124	-0.345
w_{a1}	0.298	0.187	0.126	0.277
w_{c0}	-0.102	-0.014	-0.092	-0.705
w_{c1}	0.082	0.016	0.092	0.459

表 4.2: 多項式係数ベクトル (分類開始時)

	小説	新聞	特許	話し言葉
w_{v0}	-0.563	-0.782	-0.767	0.085
w_{v1}	0.616	0.795	0.765	0.258
w_{a0}	-0.347	-0.192	-0.145	-0.316
w_{a1}	0.308	0.18	0.143	0.269
w_{c0}	-0.09	-0.026	-0.088	-0.769
w_{c1}	0.076	0.025	0.092	0.474

表 4.3: 多項式係数ベクトル (分類終了時)

4.4 結論

本稿では，品詞分布による文書ストリームの分類法を提案した．ジャンルごとのガウス分布のパラメータと多項式係数ベクトルを更新することにより高い精度で分類が行えることを示した．

第5章　トピックモデルのラベリングによる潜在的コミュニティの分析

近年、トピックモデルを用いた研究が多方面で提案されている。例えば、ソーシャルネットワークのユーザ分析やコミュニティ抽出に用いる。しかしながら、ここで言うトピックとは、確率的基準による一種のクラスタであり、ユーザやコミュニティの意味を明示的に捉えることは自明ではない。本研究では、予め与えたラベルを用いて潜在的トピックをラベルつけし、明示的な意味を付与する手法を提案する。

5.1　前書き

近年，トピックモデルは文書集合の分析に非常に有効なモデルとして注目されており，ジャンル分類や文書要約など幅広く用いられてる [15]．また，twitter や facebook に代表されるソーシャルネットワークは極めて重要な情報源として注目されており，トピックモデルによるソーシャルネットワークのユーザ分類やコミュニティ抽出が行われている [24][25].

トピックモデルとは Blei らによって提案された潜在的ディリクレ配分法に代表される確率的生成モデルであり，対象の文書や抽出したい知識に合わせてパラメータを追加することで，柔軟にモデルを構築することができる．例えばソーシャルネットワークを対象としたモデル化では文書の著者，文書の受信者，コミュニティ等のパラメータを加えることで，文書の内容とリンク構造を加味して潜在的コミュニティを抽出できることが示されている [19][20]．これらのモデルでは著者とコミュニティがトピックによって特徴付けられているが，トピックとは確率的基準によって集められた一種のクラスタであり，潜在的トピックの混合で表されるユーザやコミュニティの意味は自明ではない。また，トピックには中身に一貫性のあるトピックと一貫性が定かでないトピックが混在する．この一貫性が無いトピックはストップワードのような多数の文書で頻度が高くなる語が集まっていることが多く，出現回数が多いことからトピック分布の中でも確率が高くなる傾向にあることから，トピック分布の解釈が困難となる．このためトピック分布の混合で表される潜在的コミュニティに自動的にラベル付けする手法が望まれる．

トピックのラベル付けを行う手法はいくつか提案されており，Mei らにトピックのラベル付け [21] では, 実験データ中から候補ラベルを生成し，複数のスコア値を用い

ることで，トピックのラベル付けが行えることを示している．また，Ramage らによる Labeled LDA[22] では教師有り学習を行い，ラベルにトピックを 1 対 1 で対応付けて学習することにより文書をラベルとトピックの混合としてモデル化することができ，Twitter 上のユーザやツイートの特徴付けにも用いられている [23].

これらの手法はトピックにラベル付けする手法としては有用であるが．ラベルとトピックの関係しか考慮していないため，ユーザ分布とトピック分布から抽出される潜在的コミュニティのラベル付けに用いるには不十分である．例えばオリンピックや原発事故のような話題が大きく偏るイベントが発生した際に，コミュニティごとに特徴があるにも関わらず，どのコミュニティでも特定のトピックの確率が高くなり、一部のトピックの影響により各コミュニティで同じラベルが付けられる可能性がある．このためトピック分布のみによる潜在的コミュニティのラベル付けは困難であり，ユーザ分布やリンク構造など他の情報も加味する必要がある．

本研究では、実験データから抽出された候補ラベルを用いて潜在的トピックをラベル付けし、トピック分布とユーザ分布を加味してコミュニティにラベル付けすることで，潜在的コミュニティに明示的な意味を付与する手法を提案する．

第 2 章では関連研究について述べ，第 3 章ではコミュニティ抽出手法について述べる．第 4 章ではラベリング手法について述べ．第 5 章では実験により有効性を示す．第 6 章で結論とする．

5.2 関連研究

5.2.1 トピックモデル

トピックモデルとは, 1 つの文書が複数のトピックの混合として表現されるという仮定である．1 つの文書が 1 つのトピックで表される混合多項分布に比べ, トピックモデルは文書が複数のトピックの混合分布として, 各トピックが単語の分布として表現され, 高い精度で文書をモデル化する可能性がある.

ソーシャルネットワークを対象としたモデルとしては Pathak らによる CART(Community-Author-Recipient-Topic) モデルが挙げられる．CART モデルではネットワークのリンク構造と文書の内容の両方を考慮してコミュニティを抽出することができる．トピックがコミュニティ，著者，受信者によって決まるとし，各パラメータを加えることでモデル化を行っている．ここで CART モデルのグラフィカルモデルを図 5.1 に示す．パラメータとして α,β,γ はディリクレ事前分布のパラメータ，θ はコミュニティ空間の多項分布,ψ はトピック空間の多項分布，λ は単語空間の多項分布である．ρ は受信者のセットであり，文書には複数の受信者が含まれる可能性を考慮している．この受信者のセットから 1 単語ごとランダムに受信者が選ばれ，コミュニティ c, 著者 a, 受信者 r の関係からトピックが決まる．これによりコミュニティごとに異なる著者の分布とトピックの分布が得られる．このモデルはソーシャルネットワークからコミュニティ抽出

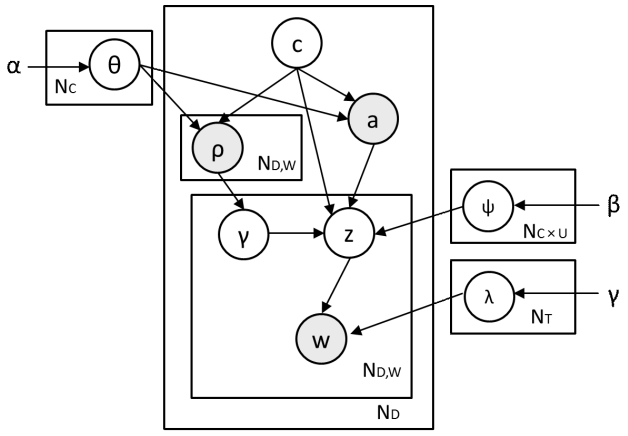

図 5.1: CART モデル

するモデルであるが，比較的小規模な会社内での E メールネットワークである Enron email コーパスを対象としており，大規模なソーシャルネットワークからコミュニティ抽出は行っていない．

5.3 コミュニティ抽出

5.3.1 Twitter からのコミュニティ抽出

本章では、トピックモデルを用いて Twitter の特徴を考慮した潜在的なコミュニティ抽出手法を提案する．大規模なソーシャルネットワークである Twitter に対応するために，CART モデルを変形しコミュニティ抽出に用いる．

Twitter 上では多数のユーザが文章を投稿しており，これらは特定のユーザへの投稿であるリプライや，他のユーザの投稿を再投稿するリツイートによってユーザ同士が繋がっている．Twitter からのコミュニティ抽出をモデル化するにあたり，ツイートの送り手と受け手のようなネットワーク構造を考慮することは重要である．しかしながら，トピックがコミュニティ, 著者，受信者の関係から求まるとした場合，必要なパラメータ数はコミュニティ数，著者数，受信者数の組み合わせであり，非常に多数のユーザが存在する Twitter では受信者をパラメータ化することは困難である．また、1 つのツイートは 140 文字以内の文字列であり，多くのツイートには数個の単語しか使われていない．このような Twitter の特性に対応するため，本手法では受信者のパラメータをモデルに組み込まない代わりに文書整形を行う．一定時間内ではユーザは似たトピックに関してツイートすると仮定し，同一時間内で同じユーザのツイートを受信者ごとに 1 つにまとめる．あるユーザが時間 T で複数回のツイートを行った場合，時間 T 内でのリプライの付いていない文書のまとまり 1 つとリプライしたユーザの種類数

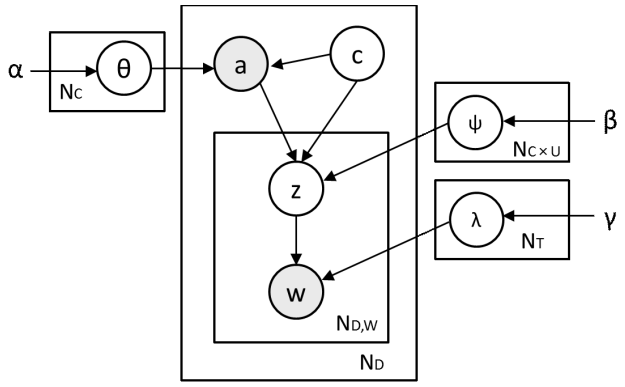

図 5.2: CAT モデル

の文書を持つ．これにより文書の単語数を増やすことができ，さらに受信者の違いによる話題の違いを考慮することが可能となる．

ツイートから単語を抽出するには文字列を分割する必要があるが，ツイートでは新聞や小説などの文章より口語に近いくだけた表現が多く使われているため，形態素解析の精度が著しく悪化することが知られている．このため、本研究では文章から 2 文字以上連続する漢字、カタカナのみを抽出して単語として用いる．また，連続する文字数が多くなると複数の単語がまとまって１つの語として抽出されていることが考えられるので，抽出された語に対して形態素解析を行い単語を分割する．例えば、”昨日、首相官邸前抗議行動に行ってきた”という文章からは”昨日”,”首相官邸抗議行動”という 2 つの単語が抽出される。さらに各語に形態素解析を行うことで”昨夜”,”首相”,”官邸”,”前”,”抗議”,”行動”に分解され,2 文字以上の単語のみを用いることから，最終的に”昨日”,”首相”,”官邸”,”抗議”,”行動”の 5 つがこの文章の単語として抽出される．

5.3.2 コミュニティ抽出手法

潜在的コミュニティの抽出には文書が属するコミュニティと著者の関係をモデル化した CAT(Community-Author-Topic) モデルを用いる．グラフィカルモデルを図 5.2 に示す．パラメータとして α,β,γ はディリクレ事前分布のパラメータ，θ はコミュニティ空間の多項分布,ψ はトピック空間の多項分布，λ は単語空間の多項分布である．この生成モデルの生成過程は以下の通りである．

1. 文書 d を生成するために，コミュニティ c_d が一様に選ばれる．
2. コミュニティ c_d を基に, 著者 a_d が選ばれる．
3. コミュニティ c_d と著者 a_d を基に，トピック $z_{(d,i)}$ が選ばれる．
4. トピック $z_{(d,i)}$ を基に，単語 $w_{(d,i)}$ が生成される．

このモデルでは著者はコミュニティに依存しており，トピックはコミュニティと著者に依存して選ばれる．これによりコミュニティに所属している著者の分布とコミュニティから生成されたトピックの分布が得られる．これらの結合確率は以下の式で表される．

$$
\begin{aligned}
&p(c_d, a_d, \mathbf{z}_d, \mathbf{w}_d) \\
&\quad = p(c_d) pr(a_d|c_d) \prod_{i=1}^{N_d} p(w_{(d,i)|z_{d,i}}) p(z_{(d,i)}|c_d, a_d)
\end{aligned} \tag{5.1}
$$

潜在変数の推定する手法としては，EM アルゴリズムや変分ベイズ法が挙げられるが，本研究では十分な反復回数の基で高い精度で潜在変数を推定できるギブスサンプリングを用いる．コミュニティに関する更新式は以下の式で求まる．

$$
\begin{aligned}
&p(c_d = c|\mathbf{c}_{-d}, \mathbf{a}, \mathbf{z}, \mathbf{w}) \\
&\propto \frac{(n^{CU}_{-d,cu_i} + \alpha)}{\sum_{u=1}^{U}(n^{CU}_{-d,cu_i}) + U\alpha} \\
&\times \frac{\prod_{z=1}^{T}\Gamma(e_{d,z} + n^{(CU)T}_{-d,(c_d a_d)z}) + \beta)}{\Gamma(\sum_{z=1}^{T}(e_{d,z} + n^{(CU)T}_{-d,(c_d a_d)z})) + T\beta}
\end{aligned} \tag{5.2}
$$

ここで n^{CU}_{-d,cu_i} は文書 d を除いてコミュニティ c からユーザ u_i が生成された回数,$e_{d,z}$ は文書 d でトピック z が生成された回数,$n^{(CU)T}_{-d,(c_d a_d)z}$ は文書 d を除いてコミュニティ c_d, 著者 a_d の組み合わせでトピック z が生成された回数,U はユーザの数,T はトピックの数である．次に各単語に対して割り振られるトピックに関する更新式は以下の式で求まる．

$$
\begin{aligned}
&p(z_{d,i} = z|\mathbf{c}_{-d}, \mathbf{a}, \mathbf{z}_{-(d,i)}, \mathbf{w}_{-(d,i)}, c_d, a_d, w_{d,i} = w) \\
&\propto \frac{n^{TW}_{-(d,i),zw} + \gamma}{\sum_{v=1}^{W} n^{TW}_{-(d,i),zv} + W\gamma} \times \frac{n^{(CU)T}_{-d,(c_d a_d)z} + \beta}{\sum_{z=1}^{T} n^{(CU)T}_{-d,(c_d a_d)z} + T\beta}
\end{aligned} \tag{5.3}
$$

ここで $n^{TW}_{-(d,i),zw}$ は文書 d の i 番目の単語以外で単語 w にトピック z が割り振られた回数，V は総単語数である．

5.4 ラベリング手法

5.4.1 候補ラベルの生成

本研究では，実験データ中から抽出された候補ラベルを用いてコミュニティのラベル付けを行う．コミュニティの内容を表す語の単位として 1 単語では不十分であるた

め，候補ラベルの語の単位を 2-gram とし，実験データの各文書から 2-gram を抽出して出現頻度の上位 1000 語を候補ラベルとする．2-gram は繋がりあった語を抽出するのに有用であるが，ここで抽出される候補ラベルは単語頻度に影響を受けるため，各単語の頻度が高い語のみとなる．例えば，頻度の高い”今日”や”明日”などに続く語は繋がりが弱い語であっても多くの語が候補ラベルとなる．逆に”太陽光”と”発電”という語の間には強い繋がりがあると考えられるが，各単語自体の頻度が高くないために候補ラベルとならない．このため高頻度語以外の候補ラベルを抽出するために，語の繋がり捉える指標としてコロケーション抽出などに用いられる相関ルールの確信度を使用する．共起する語 w_1,w_2 の出現回数を n_1,n_2, 共起回数を n_{12} としたとき，w_1 を条件として w_2 が出現する確信度 $w_1 \Rightarrow w_2$ 以下の式で求まる．

$$C(w_1 \Rightarrow w_2) = \frac{n_{12}}{n_1} \tag{5.4}$$

ここで求まる確信度の上位 1000 語を候補ラベルに加える．

5.4.2 ラベルの推定

まず潜在的コミュニティのラベル付けをするために潜在的トピックのラベル付けを行う．トピックはそれぞれ単語分布を持っていることから，トピック z の基でラベル l が出現する確率を求め，尤度の最も高い候補ラベルをトピックのラベルとする．トピック z のラベル l は以下の式で求まる.

$$Ans(l) = \arg\max_{l} p(l|z) \tag{5.5}$$

次に潜在的コミュニティのラベル付けを行う．潜在的コミュニティはトピックの確率分布 T とユーザの確率分布 U で表されており, この 2 つの確率分布を掛け合わせることで候補ラベルを決定する．これにより，例えトピック分布が同じであってもユーザ分布の違いにより異なるラベルが付けられる．コミュニティのラベルとなる候補ラベル l は以下の式より求まる．

$$Ans(l) = \arg\max_{l} p(l|z)p(z|u,c) \tag{5.6}$$

5.5 実験

本章では 2 つの実験を行う．第 1 に Twitter から潜在的コミュニティの抽出を行う．第 2 に抽出された潜在的コミュニティに対してラベル付けを行う．

ユーザ数	文書数	異語数	総単語数
8870	169979	63778	3166073

表 5.1: 実験データ

5.5.1 実験準備

実験に用いる Twitter のコーパスは 2012 年 7 月 1 日の 1 日分であり，この中でツイート数が上位 10000 となるユーザを使用する．ツイート数が非常に多いユーザに関しては”bot”[1]である可能性が高くなるので，1 日で 200 回以上のツイートを行っているユーザを除外する．3 章で述べた手法によりツイートから語を抽出し，抽出された語の形態素解析には mecab を用いる．2 時間ごとにツイート情報をまとめる．ここで文書の単語数が 2 以下である文書では正確なトピックを推定することが困難なため取り除く．また，実験データ中でノイズとなる出現頻度が 1 の単語と高頻度語となりトピックの語分布の特徴付けを妨げる出現頻度が 4000 以上の単語を取り除く．取り除かれる単語の異語数はそれぞれ 66026 語，87 語である．最終的に使用する実験データのユーザ数，文書数，異語数，総単語数を 5.1 に示す．相関ルールの最小支持度は 300/総単語数 とし，トピックモデルの各パラメータは $\alpha = 0.01$,$\beta = 0.01$,$\gamma = 0.01$, トピック数 40, コミュニティ数 6, ギブスサンプリングの繰り返し回数 200 とする．

5.5.2 比較方法

抽出された確率分布の比較には JS ダイバージェンスを用いる．JS ダイバージェンスとは左右対称な確率分布の差の尺度であり確率分布 P と確率分布 Q が与えられたとき，JS ダイバージェンスは以下の式で求まる．

$$JS(P//Q) = \frac{1}{2}(\sum_x P(x)log\frac{P(x)}{R_(x)} + \sum_x Q(x)log\frac{Q(x)}{R(x)}) \tag{5.7}$$

ここで R は確率分布 P と Q の平均であり，$R = \frac{P+Q}{2}$ となる．確率分布の差が大きいほど JS ダイバージェンスの値も大きくなり P と Q が等しいとき JS ダイバージェンスは 0 となる．

5.5.3 実験結果

第 1 の実験より，抽出された潜在的コミュニティのトピックとユーザのトップ 10 を表 5.2 に示す．各コミュニティで最も出現する確率の高いトピックはコミュニティ 1 から順

[1]Twitter 上には大量の文章を自動的に投稿するプログラムである bot が多数存在する

C1		C2		C3	
ユーザ番号	トピック	ユーザ	トピック	ユーザ番号	トピック
956	16	847	28	4145	33
471	34	1521	30	4256	16
6381	20	2738	15	67	28
7884	33	744	20	844	11
69	28	925	33	2536	20
1521	15	2173	34	2705	34
2640	30	6602	11	7884	18
5895	18	67	16	1547	30
8351	11	572	10	1970	9
1476	2	3463	18	1393	15
C4		C5		C6	
ユーザ番号	トピック	ユーザ番号	トピック	ユーザ番号	トピック
695	16	4045	16	2296	28
6732	30	6512	28	480	16
910	28	2037	18	2640	11
1970	18	4059	34	2705	36
3191	33	5530	33	4404	33
3301	34	2536	20	2536	20
5516	15	2640	10	3405	15
6582	11	7640	11	5654	30
471	9	7765	30	8816	34
801	10	553	36	1212	10

表 5.2: 各コミュニティのユーザとトピックのトップ 10

に 16,28,33,16,16,28 であり，コミュニティ1,4,6 とコミュニティ2,6 で同じトピックが最上位に出現している．各コミュニティで最上位となるユーザは 956,847,4145,695,4045,2296 と全て異なっている．次に各コミュニティでトップになったユーザのトピック分布を 5.4 に示し，トピックの単語分布を表 5.5 示す．5.4 はコミュニティで最上位になったユーザであるが，ユーザ 695,2296 以外では最上位となるトピックが異なっている．各コミュニティのユーザ分布とトピック分布を JS ダイバージェンスにより比較した結果を表 5.3 に示す．次に第 2 の実験より，トピックのラベル付けの結果を表 5.6 に示し，潜在的コミュニティのラベル付け結果を表 5.7 に示す．各トピックに付けられるラベルはすべて異なっており，同じトピックが最上位になるコミュニティであっても異なるラベルが付けられている．

5.5.4 考察

表 5.2 より，各コミュニティのトピック分布は多くの同じトピックが上位に出現しているが，ユーザ分布ではトップ 5 に含まれる著者がすべて異なっている．これは表 5.3 の JS ダイバージェンスによる比較が示すようにコミュニティ同士のトピック分布の JS ダイバージェンスの平均が 4.19×10^{-4} であるようにトピック分布の差が小さくなっているためであると考えられる．コミュニティ1,4,5 でトップになったトピック 16 は，コミュニティ2,3,6 でも上位に出現しいる．トピック 16 のラベルでは上位語同士の組み合わせのラベルが出現しておらず，3 番目のラベルでは”チケット”,”発売”の両方がトップ 10 に含まれていない．これはトピック内の上位語に対する関連した語が集まってい

	C1	C2	C3	C4	C5	C6
(トピック)						
C1		3.35×10^{-4}	3.00×10^{-4}	3.62×10^{-4}	3.55×10^{-4}	5.03×10^{-4}
C2			3.63×10^{-4}	4.88×10^{-4}	3.62×10^{-4}	3.38×10^{-4}
C3				4.16×10^{-4}	4.68×10^{-4}	5.16×10^{-4}
C4					3.71×10^{-4}	6.34×10^{-4}
C5						4.77×10^{-4}
C6						
(ユーザ)						
C1		0.0342	0.0341	0.0338	0.0338	0.0330
C2			0.0343	0.0338	0.0340	0.0333
C3				0.0339	0.0341	0.0336
C4					0.0337	0.0331
C5						0.0334
C6						

表 5.3: コミュニティ同士の比較

956		847		4145		695		4045		2296	
トピック	確率	トピック	確率	トピック	確率	トピック	確率	トピック	確率	トピック	確率
2	0.144	7	0.119	33	0.331	6	0.246	22	0.160	6	0.0976
34	0.141	35	0.112	15	0.122	5	0.136	38	0.112	10	0.0918
33	0.132	21	0.105	37	0.093	30	0.126	3	0.085	25	0.0861
22	0.100	20	0.085	35	0.086	24	0.087	9	0.082	7	0.0746
38	0.072	36	0.066	26	0.077	16	0.085	8	0.080	31	0.0660
1	0.063	0	0.063	8	0.075	15	0.053	15	0.078	4	0.0617
10	0.039	16	0.058	19	0.045	38	0.051	35	0.056	19	0.0516
28	0.039	23	0.054	13	0.039	28	0.041	0	0.051	2	0.0502
12	0.037	6	0.046	32	0.039	37	0.028	10	0.046	12	0.0473
15	0.035	32	0.044	9	0.027	34	0.026	25	0.046	11	0.0430

表 5.4: ユーザのトピックトップ 10

ないためであり，様々な文書で現れた語の集まった一貫性の無いトピックであると考えられる．表 5.5 と表 5.6 よりトピックのラベルとなる語はトピックの単語分布でトップ 10 に入る語の組み合わせだけでなく，トップ 10 に入ってない語との共起語が選ばれている．これは相関ルールの確信度による候補ラベルを加えたためだと考えられる．またトピック 2 のラベルでは実験データの日時である 7 月 1 日に公開する映画に関するラベル，トピック 4 では”プール 冷却”,”冷却 装置”など実験データ中に起こったイベントに関するラベルが付けられている．コミュニティのラベルでは，コミュニティでトップになったトピックとは違うラベルが付けられている．表 5.4 で示した各コミュニティでトップになったユーザのトピックとコミュニティで上位に出現するトピックが異なっているためであり，コミュニティのラベル付けにユーザの分布も条件に与えたためであると考えられる．

1		2		4		16		28		33	
単語	確率	単語	確率	単語	確率	単語	確率	単語	確率	単語	確率
暴力	0.00304	新宿	0.00429	冷却	0.01040	開始	0.00220	花火	0.00239	英語	0.00207
抗議	0.00258	発表	0.00272	プール	0.00783	予約	0.00163	試合	0.00228	社会	0.00193
市民	0.00207	公開	0.00265	装置	0.00554	韓国	0.00157	大会	0.00225	学校	0.00185
機動	0.00198	予約	0.00189	燃料	0.00547	大変	0.00154	風呂	0.00178	人人	0.00170
大事	0.00196	部屋	0.00182	停止	0.00536	全員	0.00153	言葉	0.00177	テレビ	0.00167
警察	0.00194	風呂	0.00160	使用	0.00486	テレビ	0.00151	ドン	0.00177	発売	0.00166
ダンナ	0.00194	開催	0.00158	東電	0.00363	結構	0.00149	大変	0.00171	結構	0.00166
大変	0.00181	大変	0.00151	事故	0.00342	開催	0.00147	韓国	0.00165	最初	0.00152
女性	0.00180	上映	0.00151	連絡	0.00260	元気	0.00146	来週	0.00158	数学	0.00150
結構	0.00176	女性	0.00148	状況	0.00253	リプ	0.00142	結構	0.00154	女性	0.00149

表 5.5: トピックの単語トップ 10

	トピック 1	トピック 2	トピック 4	トピック 16	トピック 28	トピック 33
1	抗議 活動	公開 新宿	プール 冷却	生放送 開始	花火 大会	発売 決定
2	抗議 行動	新宿 バルト	冷却 装置	チケット 予約	大会 コピー	数学 理科
3	官邸 抗議	ヱヴァ 公開	装置 冷却	チケット 発売	恋人 花火	公開 決定

表 5.6: トピックのラベリングトップ 3

5.6 結論

本論文では，潜在的コミュニティのラベリング手法を提案した．トピックモデルにより Twitter から潜在的コミュニティを抽出することができ，ラベル付けの際にユーザ分布を条件として与えることでコミュニティのトピック分布の差が少なくとも異なるラベル付けが行えることを示した．

	C1	C2	C3	C4	C5	C6
1	一方 詐欺	リフォロー 是非	配信 開始	プール 冷却	抗議 活動	男子 仕方
2	一方 拒否	パチンコ 野球	全員 全力	行動 報道	抗議 行動	カップル 会話
3	一方 会話	相撲 リフォロー	チケット 発売	冷却 装置	官邸 抗議	カップル 注

表 5.7: コミュニティのラベリングトップ 3

第6章　結論

本研究ではトピックモデルや品詞分布を用いることで，文書の著者や話題，ジャンルによらず文書分類が行えることを示した．文書分類では文書の特性によって特徴量に差があることから分類に必要な手法が異なる．また，文書ストリームやコミュニティの抽出など目的に応じて分類手法の特性を生かすことで有効な分類が行える．

まず，トピックモデルによる著者推定では，LDA の構造内で著者推定がカテゴリ独立に行えることを示した．すなわち，トピックが著者間の潜在的な区別できる特性を捉えることから，トピックモデルはカテゴリから独立していることを経験的に示した．

品詞分布による著者推定では，品詞分布の線形回帰と名詞の事前分布にガウス分布を用いるという新しい分類法を示した．これにより高い精度でジャンル分類ができ，LDA と比較して極めて短い時間で分類が行えることを示した．また，品詞分布の特性は文書のジャンルごとに異なっており，ジャンルごとの回帰直線を用いることにより品詞分布の特性が得られることを検定により示した．

品詞分布による文書ストリームの分類では，線形回帰とガウス分布の各パラメータを差分学習してパラメータを更新することにより効果的に文書ストリームの分類が行えることを示した．提案手法ではパラメータを更新しない比較手法と比べて精度が高いことから本手法は有効であるといえる．

トピックモデルのラベリングでは，トピックにラベル付けをする手法を示した．これにより文書をコミュニティに分類でき，かつ抽出したコミュニティを解釈可能なものとした．

参考文献

[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] Griffiths, T.L. and Steyvers, M.: Finding Scientific Topics, proc. National Academy of Sciences 101, 2004

[3] Hofmann, T.: Probabilistic Latent Semantic Indexing, SIGIR, 1999

[4] Holmes, D. and Forsyth, R.: The Federalist revised: New directions in authorship attribution, Literary and Linguistic Computing 10-2, pp.111-127, 1995

[5] Mendelhall,T.C.: Mechanical Solution of a LiteraryProblem, Popular Science Monthly 60-2, 1901

[6] 金明哲: 読点と書き手の個性, 計量国語, Vol18,No8,pp.382-391,1993.

[7] 松浦司，金田康正: n-gram 分布を用いた近代日本語小説文の著者推定, 自然言語処理, 134-5,1999.

[8] Nakayama, M. and Miura, T.: Identifying Topics by using Word Distribution, proc. PACRIM, 2007

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

[10] 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.

[11] 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.

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

[13] 水谷 静夫: 大野の語彙法則について計量国語学,vol35,pp.1-13,1965

[14] 金 明哲, 村上 征勝:　ランダムフォレスト法による文書の書き手の同定統計数理,vol55,pp255-268,2007

[15] Shirai, M. and Miura, T.: On Domain Independence of Author Identification,IDEAL, 2011

[16] Chai, K.M., Ng, H.T. and Chieu, H.L.: “ Bayesian Online Classifiers for Text Classification and Filtering ”, Proc.ACM (SIGIR),2002

[17] Gurmeet Singh Manku, G.S. and Motwani, R.: “ Approximate Frequency Counts over Data Stream ”, Proc. 28th international conference on Very Large Data Bases (VLDB), 2002

[18] Masato Shrai, Takao Miura. ”Document Classification Using POS Distribution ” 16th East-European Conference on Advances in Databases and Information Systems(ADBIS),2012

[19] Ding Zhou, Eren Manavoglu, Jia Li, C.LeeGiles, Hongyuan Zha“Probabilistic Models for Discovering E-Communities”,Proceedings of the 15th international conference on World Wide Web(WWW’06),2006

[20] Nishith Pathak, Colin DeLong, Kendrick Erickson,and Arindam Banerjee “Social Topic Models for Community Extraction”, The 2nd SNA-KDD Workshop’08(SNA-KDD’08), 2008.

[21] Qiaozhu Mei, Xuehua Shen, and Chengxiang Zhai “Automatic Labeling of Multinomial Topic Models”,Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining(KDD’07),pp.490-499,2007.

[22] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D.Manning “Labeled LDA:A supervised topic model for credit attribution in multi-labeled corpora”,Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing(EMNLP’09),pp.248-256,2009.

[23] Daniel Ramage, Susan Dumais, and Dan Liebling “Characterizing Microblogs with Topic Models”,AAAI’10,2010

[24] Liangjie Hong, and Brian D.Davison “Empirical Study of Topic Modeling in Twitter”,1st Workshop on Social Media Analytics(SOMA’10),2010

[25] Mrinmaya Sachan, Danish Contractor, Tanveer A.Faruquie, and L. Venkata Subramaniam “Using Content and Interactions for Discovering Communities in Social Networks”,WWW’12,2012