平成 28 年度　卒業論文

テキストマイニングによる Twitter 個人アカウントの性格推定

法政大学　理工学部　経営システム工学科

経営数理工学研究室

2017 年 2 月

学籍番号:13X4145

吉村　潤平

指導教員:五島洋行　教授

学科名	経営システム工	学籍番号	13X4145
申請者氏名		吉村　潤平	
指導教員氏名		五島　洋行	

論文要旨

論文題目	テキストマイニングによる Twitter 個人アカウントの性格推定

本研究では，Twitter の個人アカウントから性格推定をすることができるか考察する．エゴグラムを用いた性格推定においては，医師の診断やペーパーテストなどの方法が知られている．医師の診断は最も基本的な方法ではあるが，金銭的・時間的な面で問題がある．また，ペーパーテストはローコストかつ短時間で結果が得られるが，理想像を思い描くことで自分を誇張することができる．個人の正しい性格をローコストかつ短時間で推定するために，SNS を用いてはどうかと考える．そこで本研究では，Twitter を用いた性格推定手法を提案する．ナイーブベイズ法により自動分類した結果，エゴグラム診断サイトとの一致率が 48%という結果が得られた．

目次

第1章 はじめに

本論文では，Twitter の個人アカウントにおける投稿文章を分析し，性格を抽出する．これにより，正確かつローコストで性格推定ができると考えられる．

本章では，研究の背景と目的について述べたのち，先行研究を紹介する．

1.1背景と目的

近年，マイクロブログや SNS の普及で，個人の行動や感情を容易に文章で記録でき，また大量のデータとして蓄積できるようになった．特に，文章の短さによる手軽さや，情報発信のリアルタイム性といった特徴をもち，また，コミュニケーションツールとしての側面をもつ Twitter には，日常の些細な記録や企業広告，地震速報など様々なデータが混在している．こういった大量のテキストデータから，個人の興味・嗜好，場所，日時など様々な情報を抽出することで，企業のマーケティング活動や株価予測，選挙結果予測に活用するテキストマイニングという手法が注目されている．テキストマイニングとは，自由に記述された膨大なテキストデータから目的に合った情報を抽出し分析する手法のことである．

マイクロブログや SNS の個人のテキストデータは，文章の書き手によって性格が表れるという側面をもつと考えられる．例えば，全く同じ体験をした人が，それを全く同じ文章で表現する可能性はきわめて低い．そのときに感じたうれしい，かなしいなどの感情や考えが文章の単語や句読点の位置などによって現れる．すなわち，マイクロブログや SNS の個人のテキストデータを分析することでその個人の性格を推定することができると考える．

心理学の分野で，エゴグラム[1]という人の内面的特徴を表現する方法があり，南川ら[2]の研究では，これを個人 Blog で使用される語句から推定することを目的とし，ナイーブベイズ法によって推定する手法を提案している．評価方法によって結果に違いは見られるが，60%から 80%の推定精度を得ている．

エゴグラムを用いた性格推定では，代表的なエゴグラムの抽出方法として主に以下の 2 種類がある．

・観察による診断

・ペーパーテスト

観察による診断は，最も基本的な方法で，専門家と被験者のカウンセリングによって行われる．被験者の発話，ジェスチャー，姿勢，表情などの視覚や聴覚から得られる情報をもとに，エゴグラムを抽出する．この方法による性格推

定は，専門家の腕によるが，一般的に性格で精密な診断結果となり，個人の治療やカウンセリングには向くといえる．しかし，企業の採用活動などの大多数が被験者の対象となる場合や自分自身をより深く知りたいといった，治療やカウンセリングの意味合いが薄れる場合，時間や費用がかかるため向かないといえる．

ペーパーテストは，あらかじめ決められた数十問の質問に被験者が回答し，その回答結果からエゴグラムを抽出する．観察の診断による抽出方法よりも，ローコストで結果を取得することができる．しかし，理想像を思い描くことで自分を誇張することができてしまい，正確な診断結果が得られにくいといえる．例えば，企業の採用活動で誇張された診断結果で採用に結びついてしまった場合に悪影響となってしまう恐れがある．

そこで本研究では，Twitter を用いた新たな抽出方法を提案する．エゴグラムによる性格診断[3]というエゴグラム診断サイトがあり，Twitter 上にこの診断サイトの結果を投稿している個人アカウントを対象に，投稿文章を収集し，ナイーブベイズを用いて自動分類する．その自動分類結果とエゴグラム診断サイトの結果の一致率を検証することが本研究の目的である．

1.2 Twitter の選定理由

Twitter とは，140 文字以内の短文を作成し，そのときの心情や行動を画像や位置情報などと共に投稿できる SNS である．Twitter を用いる理由は主に三つある．

(1) データ収集が容易である：Twitter の国内利用者数は図 1 に示すように，1,400 万人(2012 年 3 月時点)におよび，個人の行動記録として広く普及していることからデータ収集が容易に行える．

(2) エゴグラムとの相性がよい：Twitter は日々の記録を自身の言葉で表現しているため，被験者との発話や表情などの視覚情報を抽出対象とするエゴグラムと親和性が高い得しょうが抽出できると考えられる．

(3) 匿名性がある：図 2 に示すように，Facebook と比較して匿名性があるため，自分の感情に素直な表現がされる可能性が高く，正確な分類結果が得られやすいと考える．

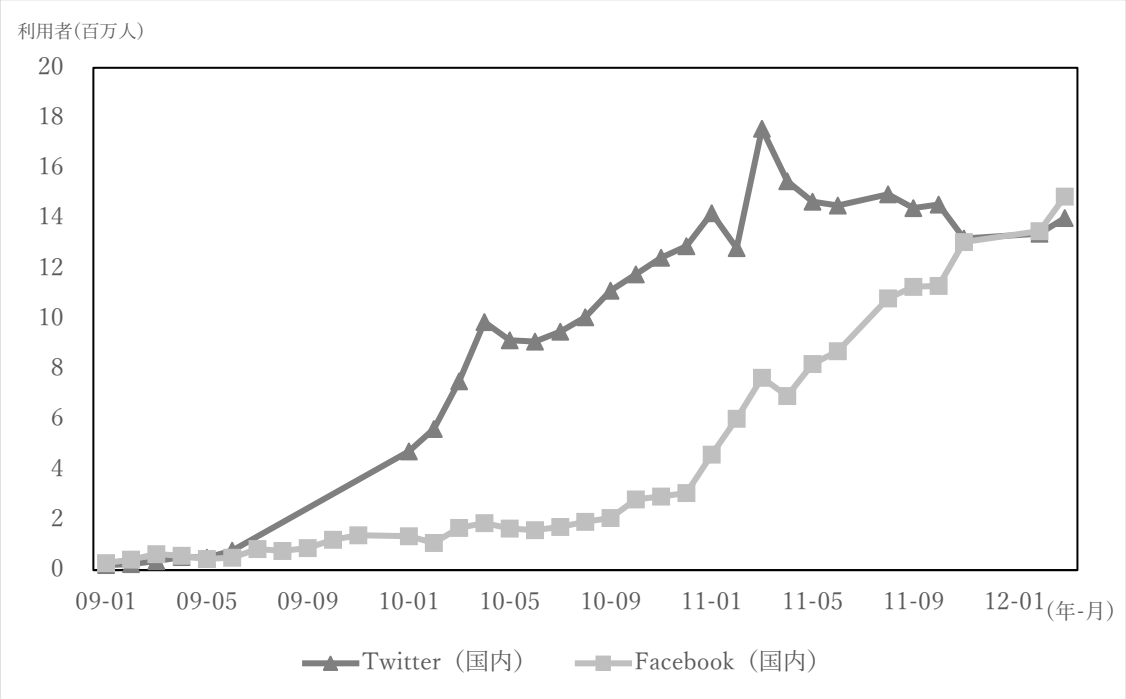

図1.　ソーシャルメディア利用者数の推移(Facebook，Twitter の例)

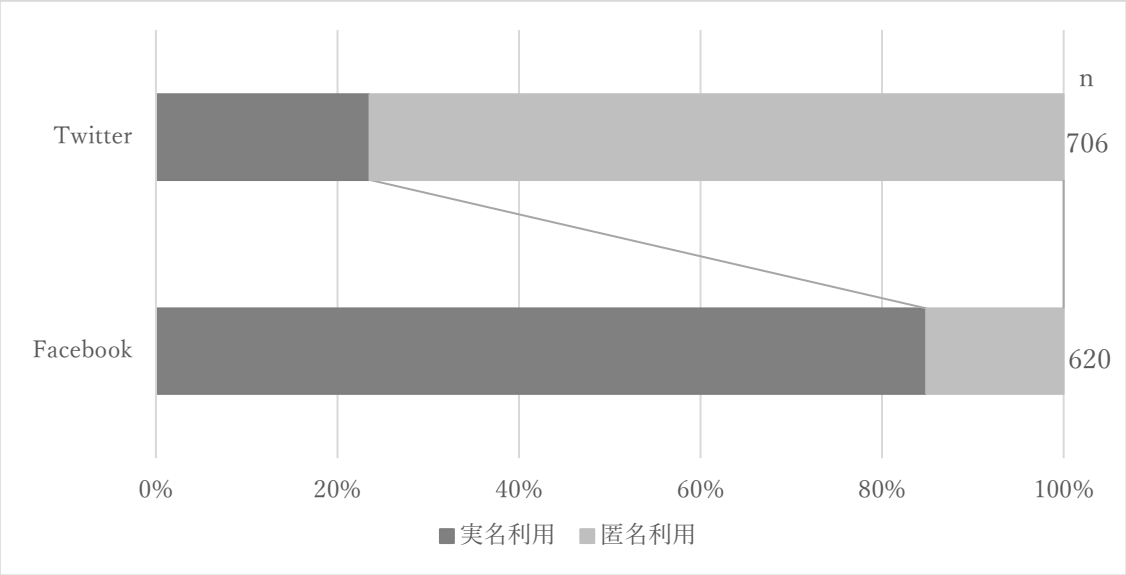

図2.　Twitter と Facebook の実名利用率

1.3先行研究

南川ら[2]の研究では，テキストマイニングによって個人 Blog データから性格を推定している．Blog の執筆者が記述したテキストデータを形態素解析し，得られた単語群を bag-of-words とみなして，ナイーブベイズ法やサポートベクタマシンを用いて分類を行う．実験データとして，551 人の Blog 執筆者から Blog データを収集し，さらに当 Blog 執筆者にエゴグラムのペーパーテストを受診してもらった回答結果を使用する．推定精度の評価は，この実験データを 10 分割交差検定によって行う．結果として 60%から 80%の推定精度となっている．以上のことから，Twitter から取得したデータを形態素解析し，ナイーブベイズ法を用いて自動分類することで性格の推定が可能であると考える．

第2章 関連知識

本章では，エゴグラムという性格推定手法について紹介する．また，分類方法として本研究で用いるナイーブベイズについて説明する．

2.1エゴグラム

エゴグラム[1]は精神科医エリック・バーンの交流分析における自我状態をもとに，弟子であるジョン・M・デュセイが考案した性格推定手法である．人の心を五つに分類し，被験者の内面的特性をグラフで表す．五つの自我状態とは，批判的親 CP(Critical Parent)，養育的親 NP(Nurturing Parent)，大人 A(Adult)，自由な子供 FC(Free Child)，従順な子供 AC(Adapted Child)である．各自我状態があらわす役割を表 1 に示す．これら五つの自我状態の相対的関係により被験者の性格を分類することができる．例えば，CP が高いと自分の価値観を正しいものと信じて譲らず，責任を持って行動し，他人に批判的であるということがいえる．

本研究で使用するエゴグラムによる性格診断という診断サイトでは，各自我状態を a，b，c の三段階評価し，ペーパーテストによって算出された結果を全 243 通りの性格分類パターンの中から抽出する．エゴグラム診断サイトによる抽出例を図 3 に示す．

表1.　エゴグラムの五つの自我状態

	Positive	Negative
CP	正義感，責任感，義務感	批判的，保守的
NP	慈愛，寛容	過保護，過干渉
A	理性的，客観的	冷徹
FC	自由奔放，創造的	自己中心的
AC	従順，謙虚，協調性	反抗的，自虐的

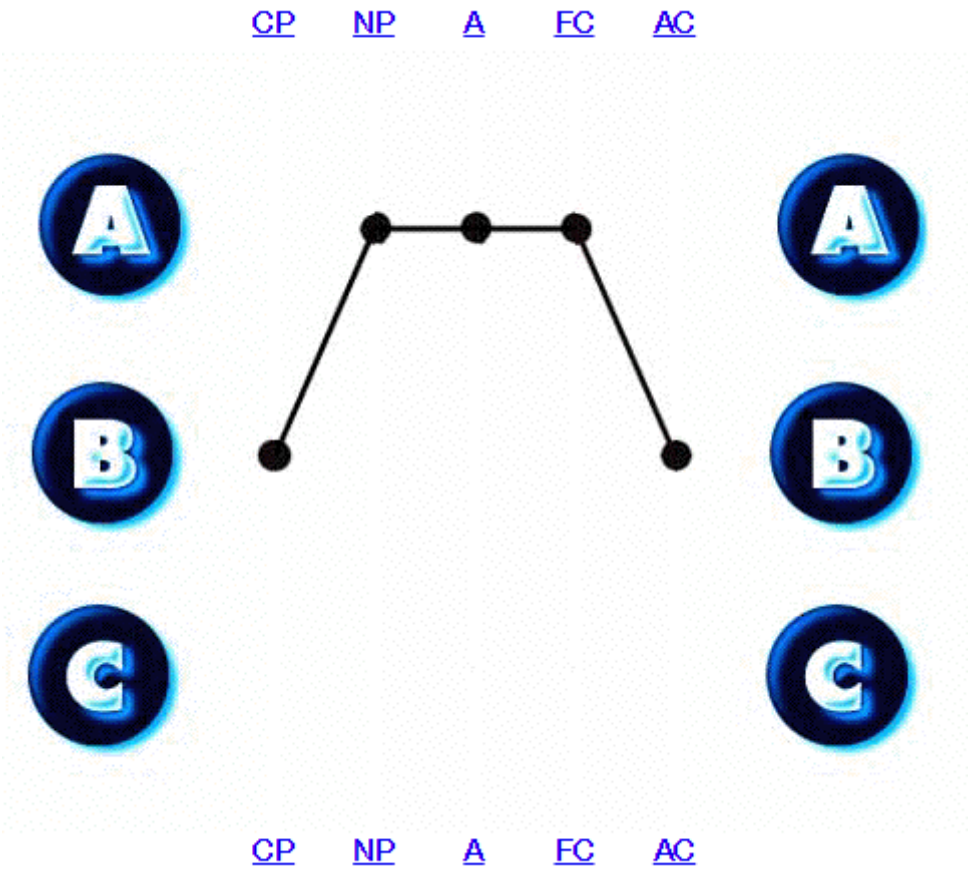

図3.　エゴグラム診断サイトによる診断結果

2.2ナイーブベイズ法

ナイーブベイズ法とは，ベイズの定理を用いた教師あり学習のテキスト分類手法である．テキスト分類とは，与えられた文書をあらかじめ与えられたいくつかのカテゴリに分類することをいう．例えば，Web ページを「政治・経済」，「科学・学問」，「コンピュータ・IT」などのカテゴリに自動分類したりする際にはテキスト分類が用いられており，すでに身近で実用化されている．これらは人間が Web ページを読み，分類しているのではなく，テキスト分類の手法を用いた分類プログラム（分類器と呼ぶ）が自動的に分類している．

分類器は，人間があらかじめ与える（文書，正解カテゴリ）を組とするデータ（訓練データと呼ぶ）をもとに各カテゴリの文書の特徴を学習する．例えば，分類器に，「iPhone」，「Apple」，「iPad」が含まれる文書は「IT」カテゴリである確率が高く，「安倍晋三」「自民党」などが含まれる文書は「政治」カテゴリである確率が高いといったカテゴリの特徴を学習させる．このように訓練された分類器を用いて，カテゴリがわからない未知の文書を自動分類させる．一般的に訓練データは多ければ多いほど分類器は正確な分類を行うことができる．以上のような人間が正解カテゴリを与えるテキスト分類手法は，教師あり学習と呼ぶ.

テキスト分類は非常に多くのアルゴリズムがあるが，本研究では，実装が簡単で，しかも高速というナイーブベイズ法を用いる．精度評価のベースラインとしてよく使われている手法の一つである．

2.3正規表現

正規表現[4]とは，文字列の集合を一つの文字列で表現するための方法である．この表現方法を利用することでたくさんの文章の中から，容易に見つけたい文字列を検索できたり，置換できたりする．正規表現は，メタ文字という特殊文字を組み合わせることで目的に沿った処理をする．表 2 にメタ文字の一部を示す.

本研究では，ツイートを取得する際に正規表現を利用する．リツイートや画像のアドレスなど意図しない文字列を取得してしまう場合に，正規表現を用いることでこれを回避する.

表2.　正規表現のメタ文字の例

.	任意の 1 文字
*	直前のパターンの 0 回以上繰り返し
+	直前のパターンの 1 回以上繰り返し
?	直前のパターンの 0 ～ 1 回繰り返し
~|~	|の左右文字列のいずれか
[~]	～のいずれか 1 文字
^	論理行頭

2.4形態素解析

言語学で，意味を担う最小の言語要素を形態素と呼ぶ．また，これに対応して自然言語処理では，文を単語ごとに分解し，どの語形変化なのかを解析する処理を形態素解析[4]と呼ぶ．

本研究では，分類器の訓練を行う際や未知の文書を分類する際，与えられた文書に対して形態素解析を行う．文書は集合として扱うため，単語の並び順は考慮されない．このようなテキスト表現は bag-of-words と呼ばれる．与えられた文書を bag-of-words とするために形態素解析を行い，文書を扱いやすくする．形態素解析を行うにあたって，本研究では，MeCab （和布蕪）[5]と呼ばれる京都大学情報学研究科−日本電信電話株式会社コミュニケーション科学基礎研究所の共同研究ユニットプロジェクトを通じて開発されたオープンソースである形態素解析エンジンを用いる．

第3章 分析手順

本章では，分析手順について述べる．Twitter 個人アカウントのデータ収集についてと収集したデータをナイーブベイズ法によって分類する計算方法を説明する．

3.1データ収集

本研究では，TwitterAPI を用いて個人アカウントにおけるつぶやきを取得する．TwitterAPI とは，Twitter 社が提供しているサービスで，Twitter の機能を呼び出すことができ，ツイートの参照や検索を行うことができる．

まず，エゴグラムによる性格診断サイトの診断結果を Twitter 上に投稿している個人アカウントを対象に，アカウント名を収集する．この結果，227 通りの性格パターンのアカウント名が収集できた．

エゴグラムによる性格診断サイトの診断結果を Twitter 上に投稿しているアカウントがなかった性格パターンが 16 通りあるが，これはそれらの性格パターンの出現頻度が低いためと考えられる．図 4 に示すように，エゴグラムによる性格診断サイトには，243 タイプ別分布状況というデータが公開されており，243 通りの性格の出現率がわかる．例えば，今回取得できなかった acccc という性格パターンがあるが，243 タイプ別分布状況のデータによると，acccc は 241 番目の出現率であり，出現数は全 10,1695 件中 1 件と非常に低い出現頻度である．本研究では，こういった出現率が低く取得できなかった性格パターンを無視し，データが収集できた全 227 通りで考察する．

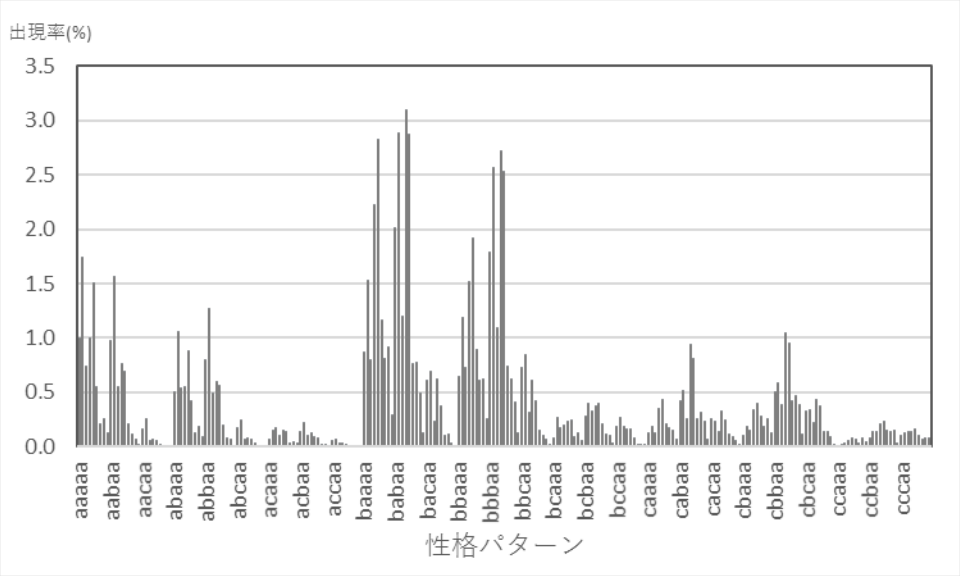

図4.　エゴグラム診断サイトによる診断結果の出現率

次に, TwitterAPI を用いて収集できた個人アカウントのツイートをそれぞれ最新 300 件取得する．取得したツイートの保管は MySQL[7]を利用する．MySQL とは，オープンソースのリレーショナルデータベース管理システム（RDBMS）で，オープンソース系としては世界で最も多く利用されている．

TwitterAPI を用いて取得したツイートはツイートをした本人が作成していない文字列が含まれてしまう場合がある．このような文字列は正規表現による置換をした後に MySQL に保管する. 正規表現による置換を用いた処理は以下の三つである．

(1) リツイートの排除
(2) ＠から始まるアカウント名の排除
(3) 画像のアドレスの排除

一つ目は，リツイートの排除である．リツイートとは他の人のアカウントのツイートを自分のツイートとして再びツイートするという Twitter の機能である．

二つ目は，＠から始まるアカウント名の排除である．Twitter の機能にはリプライという他の人のツイートに対するコメント機能がある．リプライには，誰に対するリプライかを明確にするために，全てのアカウントがそれぞれ独自に持っている＠から始まるアカウント名が含まれている．

三つ目は，画像のアドレスの排除である．ツイートの投稿は文字の他に位置情報や画像を含めることができる．画像を含んだツイートをテキストデータとして取得した場合，画像はアドレスとして取得される．

これらはツイートした本人が作成した文字列ではないため，正規表現による置換を行い削除する．この処理の目的としては，ナイーブベイズ法を用いた自動分類による結果の推定精度を向上させるためである．ツイートした本人が作成した文章とは関係のない文字列が含まれている場合，推定精度に支障をきたす可能性がある．これら三つの正規表現による置換を用いた処理の一覧と置換前・置換後の例を表 3 に示す．

表3.　正規表現による特殊文字列の置換の一覧と例

	正規表現	置換前（例）	置換後（例）
(1)	^(RT) @[0-9a-zA-Z_]{1,15}: .*	RT: 明日雪降るらしいよ	”” (空文字列)
(2)	@[0-9a-zA-Z_]{1,15}:?	@Ha_ri 楽しみ〜〜	楽しみ〜〜
(3)	(https?|ftp)(:¥/¥/[-_.!~*¥'()a-zA-Z0-9;¥/?:¥@&=+¥$,%#]+)	ハリネズミかわいい https://pbs.twimg.com	ハリネズミかわいい

3.2ナイーブベイズ法を用いた自動分類

文書ベクトルを doc = (doc_1,doc_2,…,doc_n)，カテゴリベクトルを cat とするとき，ナイーブベイズ法の中心となる式はベイズの定理を応用した式(1)で表せる．

$$P(\mathrm{cat}|\mathrm{doc}) = \frac{P(\mathrm{cat})P(\mathrm{doc}|\mathrm{cat})}{P(\mathrm{doc})} \propto P(\mathrm{cat})P(\mathrm{doc}|\mathrm{cat}) \ . \qquad (1)$$

事後確率 *P*(cat|doc)は文書 doc が与えられたとき，カテゴリ cat となる確率である．カテゴリを予測したい未知の文書は，事後確率が最も高いカテゴリへ分類する．この確率を計算するためには，事前確率 *P*(cat)と尤度 *P*(doc|cat)が必要になる．

P(doc)は文章が生起する確率だが，全てのカテゴリにおいて同じ値であるため，分類結果の算出過程では省略する．

P(cat)は，訓練データの各カテゴリの文書数の総文書に占める割合である．例えば，政治カテゴリの文書数が 70，科学カテゴリの文書数が 30，総文書数が 100 となる訓練データがあるとき，*P*(cat)は政治カテゴリが 0.7，科学カテゴリが 0.3 となる．

P(doc|cat)は，カテゴリ cat が与えられたときに文書 doc が生成される確率である．文書 doc は単語の集合[$word_1$,$word_2$,⋯,$word_k$]として表され，単語間の独立性を仮定すると式(2)で計算できる．

$$P(\mathrm{doc}|\mathrm{cat}) = P(\mathrm{word}_1, \mathrm{word}_2, \dots, \mathrm{word}_k|\mathrm{cat}) = \prod_i P(\mathrm{word}_i|\mathrm{cat}) \ . \qquad (2)$$

本来，単語の出現に独立性は成り立たない．例えば，「卒業」と「論文」は共起しやく，「先行」と「研究」は共起しやすい．これを無視し，単語の出現は独立と無理やり仮定する．このように単語間の独立性を無理矢理仮定し，文書の確率を単語の確率の積で表して単純化する点がナイーブベイズ法の特徴である．

次に，$P(\mathrm{word_i}|\mathrm{cat})$の確率が必要になる．これは，単語の条件付き確率で，カテゴリの中でその単語がどれくらいでてきやすいかを表す．訓練データのカテゴリ cat に単語 $\mathrm{word_k}$ がでてきた回数をカテゴリ cat の全単語数で割ることで算出する．T(cat,word)をカテゴリ cat に単語 word がでてきた回数，V を cat の全単語集合とすると，式(3)で表される．

$$P(\mathrm{word_i}|\mathrm{cat}) = \frac{\mathrm{T}(\mathrm{cat}, \mathrm{word_i})}{\sum_{\mathrm{word}' \in \mathrm{V}} \mathrm{T}(\mathrm{cat}, \mathrm{word}')} \ . \tag{3}$$

以上をまとめると，最大事後確率 $P(\mathrm{cat}|\mathrm{doc})$を計算することで最終的に分類されるカテゴリ $\mathrm{cat_{map}}$ が求まる．

$$\mathrm{cat_{map}} = \mathrm{argmax}P(\mathrm{cat}|\mathrm{doc}) = \mathrm{argmax}[P(\mathrm{cat}) \prod_i P(\mathrm{word_i}|\mathrm{cat})] \ . \tag{4}$$

ここで，実際に $\mathrm{cat_{map}}$ を求めるとき二つの問題がある．一つ目の問題点は，$P(\mathrm{word}|\mathrm{cat})$は非常に小さな数であり，文書中には多数の単語が含まれるため，その積はさらに小さくなり，アンダーフローを起こす可能性があることである．そこで，実際には対数をとって計算する次の式を用いる．事後確率の大小関係は対数をとっても変化しない．

$$\begin{aligned} \mathrm{cat_{map}} &= \mathrm{argmax}[\log P(\mathrm{cat}|\mathrm{doc})] \\ &= \mathrm{argmax}(\log P(\mathrm{cat}) + \sum_i \log P(\mathrm{word_i}|\mathrm{cat})) \ . \end{aligned} \tag{5}$$

二つ目の問題点は，未知の文書のカテゴリを推定する際，訓練データの全単語集合に含まれない単語を一つでも含んでいると，単語の条件付き $P(\mathrm{word}|\mathrm{cat})$ は 0 となり，$P(\mathrm{word}|\mathrm{cat})$の積である $P(\mathrm{doc}|\mathrm{cat})$も 0 となることである．例えば，未知の文書が「iPhone」，「Apple」，「iPad」の単語集合であるとき，訓練データが iPhone，Apple のカテゴリであっても，未知の文書はこのカテゴリに分類されない．「iPhone」と「Apple」がでているため，一見このカテゴリに分類される可能性が高いと考えられるが，そうならなくなってしまう．

この問題を解決するために，単語の出現回数に 1 を加えるラプラススムージングという方法を用いる．訓練データに含まれていない単語が出現すると確率が低くなるが，0 にはならない．T(cat,word)をカテゴリ cat に単語 word が出てきた回数、V を訓練データ中の全単語集合（ボキャブラリ）とすると，次の式で表せる．

$$P(\mathrm{word_i}|\mathrm{cat}) = \frac{\mathrm{T}(\mathrm{cat}, \mathrm{word_i}) + 1}{\sum_{\mathrm{word'} \in \mathrm{V}} (\mathrm{T}(\mathrm{cat}, \mathrm{word'}) + 1)} \ . \qquad (6)$$

第4章 分類結果

本章では，分類結果について述べる．第 3 章の分析手順によって得られた分類結果を示し，その分類結果の精度について検証する．

4.1 自動分類結果

収集できた性格 227 パターンの最大二つの文書のうち, 一つを未分類データ, もう一方を訓練データとし，自動分類を行った結果，一致率は平均 48%となった．ここで述べる一致率とは, エゴグラム診断サイトによる診断結果と本研究による自動分類結果の一致率である．最も一致率が高い場合で 100%，最も低い場合で 0%となった．結果の一部を表 2 に示す．また，一致率の分布を図 4 に示す．

表4.　自動分類結果の例

正解	分類結果	一致率
aaaaa	acaca	60%
abaab	cbaab	80%
bbcaa	aabbc	0%
bccbb	acaca	20%
cbccb	cbccb	100%

4.2分類精度の考察

分類精度は 48%と，精度が高いとはいえない結果となった．今回分類精度が高くならなかった要因として 4 点挙げられる．

1 点目は，訓練データの数を少ないことである．一般的に，訓練データの数が多ければ多いほど分類器は正確な分類ができるようになる．今回の自動分類では，PC のスペックの問題などもあり，各カテゴリに関して 1 人分の訓練データしか用意できていない．

2 点目は，訓練データをより正確なものに置き換えることである．今回は Twitter 上から不特定多数のアカウントを取得し，それを訓練データとしたが，東京大学心療内科による東大式 エゴグラム（TEG2）[8]などを利用し，正確な性格診断データと文書があればさらに精度は向上すると考えられる．

3 点目は，ゼロ頻度問題の補正方法としてラプラススムージングを採用していることである．ラプラススムージングは出現した全単語を単に+1 するというだけで，決して精度の高い補正方法とはいえない．最大エントロピー法や潜在変数を用いるという方法があるため，他の補正方法を検証すれば分類精度が向上する可能性がある．

4 点目は，ナイーブベイズ法を採用したことである．ナイーブベイズ法以外の手法として SVM(Support Vector Machine)や CRF(Conditional Random Fields)といった分類方法があり，これらの手法を用いて検証することで分類精度が向上する可能性がある．

第5章 おわりに

本研究では，エゴグラムによる性格診断サイトの診断結果と，ナイーブベイズ法を用いた自動分類結果の一致率を検証した．Twitter 上で性格診断サイトの診断結果を投稿した個人アカウントをデータ収集し，テキストマイニングをすることで一致率 48%の結果を得ることができた．本研究でのテキストマイニングとは，収集した文書に対して形態素解析をし，得られた単語群を bag-of-words とみなしてナイーブベイズ法を用いて文書分類をすることである．

データ収集では，推定精度を向上させるために様々な工夫を施した．例えば，エゴグラムの診断サイトによる結果を，肯定的に捉えているアカウントを優先的に収集するなどの条件を設けた．また，取得したデータに正規表現による置換を行うことで，リツイートや画像のアドレスなどのツイートした本人が作成していない文字列を削除した．

ナイーブベイズによる自動分類においても，ラプラススムージングというゼロ頻度問題の補正方法を採用し，より正確に分類されるように工夫した．

分類精度はまだまだ低く，実用的とは言えないが，ライフログという点で個人の過去から性格を推定できるようになれば，様々な場所で活用できると考える．例えば，企業の採用活動では，今までの SPI のようなペーパーテストによる性格推定であると，自らの性格を誇張することができる．しかし，個人の文章から性格を推定できれば，企業の求める人物像に近しい人を採用でき，採用後にミスマッチが発生することが少なくなる．また，近年 SNS がマーケティングとして利用されることが多いが，SNS 上で商品を購入しレビューをしている人が，どんな性格なのかという観点でマーケティングできれば，世代や性別以外の新たな観点から商品企画や広報戦略を立てることができるかもしれない．

参考文献

[1] ジョン・M・デュセイ，“エゴグラム”，創元社，2000.

[2] 南川敦宣，横山浩之，“テキストマイニングによる個人 Blog データからの性格推定手法”，データマイニングと統計数理研究会第 12 回，pp.96–100，2010.

[3] エゴグラムによる性格診断，
“http://www.egogram-f.jp/seikaku/”（2017 年 1 月 20 日確認）

[4] 宮前竜也，“正規表現”，技術評論社，2006.

[5] 松本裕治，影山太郎，永田昌明，齋藤洋典，徳永健伸，“単語と辞書”，岩波書店，1997.

[6] MeCab（和布蕪），
“http://taku910.github.io/mecab/”（2017 年 1 月 20 日確認）

[7] MySQL,
“https://www-jp.mysql.com/”（2017 年 1 月 20 日確認）

[8] 東京大学医学部心療内科 TEG 研究会，“新版 TEGⅡ 解説とエゴグラム・パターン”， 金子書房.