Topic/Author Word Model を用いたトピック推定に関する研究

中山, 基 / NAKAYAMA, Motoi

(発行年 / Year)
2008-03-24
(学位授与年月日 / Date of Granted)
2008-03-24
(学位名 / Degree Name)
修士(工学)
(学位授与機関 / Degree Grantor)
法政大学 (Hosei University)

2007年度
修 士 論 文

Topic/Author Word Model を用いた トピック推定に関する研究

RESEARCH ON IDENTIFICATION OF THE TOPIC USING TOPIC/AUTHOR WORD MODEL

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

06R3124　中山 基
Motoi NAKAYAMA

目 次

第1章　序論

1.1　問題の背景

近年のインターネットの普及により, 我々は膨大な量の情報が用意に得られるようになった. しかし, その情報量の増加は驚異的で, 今現在も毎日ページの更新と共にその情報量は増加の一途を辿っている. そしてまた, その膨大な情報量は既に我々の許容範囲を大きく上回ってしまっているのが現状である. その為, 我々の膨大な情報へのアクセスを支援するための技術が求められる.

情報へのアクセスを支援する技術としては, 文書をその内容に応じて整理・分類することが考えられる. 例として, インターネット上には Yahoo のような組織的なインデックス・サイトが数多く存在する. これらのサイトでは,Web ページを内容やトピックに応じて階層型に分類しており, トピックを検索する場合に有効な手法である. しかし, このようなサイトでは人手によりページを分類し, トピックを定めており, サーチエンジンと比べた場合にどうしても規模が小さくならざるを得ない. また, 人手による分類は, その基準が主観に依存したものになってしまう問題点も存在する. その為, データ収集が継続的かつ大規模に行われるような場合では, テキストを整理・分類するために, 計算機による自動的な処理, テキストの自動分類の技術が必要となる.

テキストの自動分類は, 大きく 2 つに分けることができる. 1 つは, 与えられた文書をあらかじめ設定されているトピック（著者, 著者の作品など）のいずれかに割り当てるものであり, もう 1 つは, 類似したグループにクラスタリングすることで, いくつかのグループに分けるものである. テキストの自動分類は,2 つの文書間の類似度を計算する事で分類する事ができ, 一般的には, すべての文書が単語に関するベクトルとして表される, ベクトル空間モデル（VSM）によって対象世界を言及する. 本研究では, 自動分類のうちでも, トピックの推定に関して扱う.

また, 一般的にテキスト形式のデータは, 次元数が数万にも及ぶ高次元データである. その為, 高次元データをそのまま扱うと, 効率, 計算機容量の確保及び即応性への対応が困難となる問題点が存在する.

本研究では, テキストを整理・分類するため, トピックの推定を行う. トピックを推定する新たな手法を提案し, 更に次元縮小手法を適用する. これにより, 情報検索の効率向上に大きなヒントを与えることができる.

1.2 関連研究

1.2.1 語の取り扱い

一般的に, テキスト形式のデータでは, テキストは語の並びとして構成されるが, その語をどのように選択するかに明確な決まりは存在しない. しかし, テキスト形式のデータでは語の同義性や類犠牲から, いくつかの語が同時に生じる可能性が高い. 実際, 同義語と多義語を考慮することで, 分類の精度を向上させる研究が行われている. 複合語や共起性の強い語を考慮するかどうかは, 分析結果に大きな影響を与える.

1.2.2 著者/トピックの推定

過去一世紀以上にわたる論争の一つに著者の推定問題がある. テキストから何らかの特徴を捉えて著者の推定を行うもので, 日本のグリコ森永事件の脅迫状の分析やシェークスピア実在論が挙げられる. これと並ぶものとして存在するのが, トピックの推定であり, 何のトピック（興味ある事柄や出来事）を論じているかを推定する. トピックを推定することは, 効率よく検索するための情報の要約, トピックへの文書の自動分類とも関連がある. また, これまでの研究によると, テキストから著者よりもトピックとの関連性を論じる方が分析しやすいことが知られている.

1.3 扱う問題

1.3.1 訓練データに対する分類精度の依存

テキストの分類には経験的にテキストの分類には教師つき学習が有効であり, 様々な機械学習手法が研究されている. 教師つき学習にはあらかじめ人手により正確に分類されたデータ（訓練データ）が必要であるが, データ収集が継続的かつ大規模に行われるような場合では, 人手を使ってデータを分類することは時間と労力などのコストの大幅な増大を意味することにもなり, その分類作業自体も主観的なものになってしまう問題がある. また, 分類の精度は訓練データに依存するため, 訓練データをどう作成するかといった事も問題となる.

本研究では,EM アルゴリズムを用い, 少量の訓練データで未分類データを分類しながら, その未分類データを訓練データとして利用するステップを繰り返すことで, この問題を解決する. この際, 一旦拡大したエラーが繰り返しにより, 修復されにくく誤りを訂正することが困難であるという問題があるが, 語の共起性を考慮に入れることでこれを解決する. 実験により本手法の有効性を示す.

1.3.2 Topic/Author Word モデル

T/AW モデルとは, 同一著者の下では, 各トピックは対応する語集合の多項式分布確率で表されるという仮定である. つまり, 著者はトピックに依存する多項式分布確率により語を確率的に選択していると仮定される. これは語の分布の検証により, 文書のトピックを推定することが可能であることを意味する. 他方,AW モデルは著者がトピックに独立に著者自身の語集合の多項式分布確率で表されるという仮定である. 一般的に T/AW モデルは広く信じられているが,AW モデルにおいてはそうではない.

本研究では,T/AW モデルおよび AW モデルが実際に成り立つかどうか検証を行い, ここでは, 様々な重み設定方式を検討し,T/AW モデルを用いて分類の精度向上に寄与する方式を探る.

1.3.3 次元縮小

一般的にテキスト形式のデータは, 次元数が数万にも及ぶ高次元データであるため, 高次元データをそのまま扱うと, 効率, 計算機容量の確保及び即応性への対応が困難となり, 性能及び精度に差が生じるという問題がある. その為, 次元縮小技法を用いることで, 高次元文書ベクトルを低次元空間に射影し, 効率よく探索範囲を絞り込む必要がある.

本研究では, 次元縮小技法を適用し,Topic/Author Word モデルへの効率的で有効な処理の方法を提案する.

1.4 論文の構成

本研究では, 以上の問題について以下の構成で論じる. 第 2 章では, 語の共起性を EM アルゴリズムに融合させることで, 少ない訓練データからテキストデータを自動分類する新たな手法を提案し, 実験によりその有効性を示す. 第 3 章では, トピック語モデルの検証を行い, モデルへの効率的で有効な処理の方法を示す. 第 4 章では, 同一著者の下での語分布からのトピックを推定できるという仮説を検証する. 著者と語分布の関係についても検証を行い, これが推定に適さないことを示す. 更に, 推定方式の改善を提案し, 実験によりその有効性を示す. 第 5 章で結論とする.

1.5 発表論文

1. 中山基, 三浦孝夫: “W 共起語を考慮した EM アルゴリズムによるテキスト分類”, データ工学ワークショップ（*DEWS*）, 2006.

2. 中山基, 三浦孝夫: “単語分布からのトピック推定”, 情報処理学会 第 142 回 データベースシステム研究会および情報処理学会 第 87 回 情報学基礎研究会 , 情報処理学会 2007-DBS-142(1).

3. Motoi, N. Miura, T.: “Identifying Topics by using Word Distribution”, *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*,2007,pp.245-248.

4. 中山基, 三浦孝夫, 塩谷勇: “Topic/Author 推定方式の改善”, データ工学ワークショップ（*DEWS*）, 2008.

第2章　共起語を考慮に入れたEMアルゴリズムによるテキスト分類

2.1　前書き

近年, インターネットを介して膨大な量のテキストデータが電子的に利用可能であり, この傾向は増加をたどる一方である．こういったテキストデータの管理・検索をより高度に行うためには, テキストの分類が欠かせない．しかし, 人手によるテキスト分類は時間と労力を要するのみならず、分類作業が主観的となることから信頼性の確保も重要な問題となる。

これを解決する技術として, 経験的にテキストの分類には教師つき学習が有効であることから，決定木やベイズ分類器などの機械学習技法が研究されている [1, 4]．　教師つき学習では，あらかじめ人手により正確に分類されたデータ（訓練データ）を用いてその性質を分析し，一般的な特性を抽出するという手法がとられる．しかし, 人手による訓練データの生成は時間と労力を要するのみならず、分類作業が主観的となることから信頼性の確保も重要な問題となる。実際問題として, 特定の応用分野を限定した場合でさえ，あらかじめ分類規則（データ) を必要とするための精度の高い訓練データを確保する必要がある。

この解決策としてテキスト自動分類技術が必要とされている．しかも, 少ない訓練データから自動分類を行うことのできるEMアルゴリズムを用いることの有用性は高い。すなわち, 少量の訓練データで未分類データを分類しながら，その未分類データを訓練データとして利用するというステップを繰り返すEMアルゴリズムは, テキスト分類の手法に有効であろう．

反面,EMアルゴリズムの問題も広く知られている [2]．まず, 結果が初期値に大きく依存しがちとなる．また，繰り返し回数をあらかじめ設定できないため，収束が遅くなることが多い．同時に，収束しすぎると過学習となることも問題である．一旦拡大したエラーは修復されにく，誤りを訂正することはきわめて困難であることも重要である．この問題を改善するためには, 分類器を多重に用意し, 用いるデータの特性に応じて変化させる方法がよいとされる [5].

そこで, 本稿では, 語の共起性を EM アルゴリズムに融合させ, テキスト分類の精度向上のための新たな方式を提案する．通常, テキストデータでは語の同義性や類義性から, いくつかの語が同時に生じる可能性が高い．そのため共起語は, 特に高頻度の場合において強い相関性があると考えることができる．この共起語を分類の際の手がかり

として追加することで, テキスト分類の精度向上をはかる. また, 本稿では語の共起を文単位で考える (つまり，共起する語は複数の文に渡ってつながりを保持するとはみなさない). 一定回数以上の共起頻度を有する語に限定することで計算の効率を高め、効果的なつながりを重視する. EM アルゴリズムを用いたテキスト分類の精度を向上できることを示すため，実験により有効性と考察を示す.

第 2 章で関連研究を示し, 第 3 章で EM アルゴリズムと共起性概念をまとめる．第 4 章では本稿で提案する方式を導入し，第 5 章で実験によりその有用性を述べる．第 6 章は結論を述べる.

2.2 関連研究

文書を分類する方法としては, 上嶋らの研究がある [9]. 上嶋らは, ベイズ法での文書分類に, 同義語と多義語を考慮することによって, その分類精度を向上をさせる手法を提案している. 通常の文書分類では, 単語の持つ意味は考慮せず, 単語を単に記号的に扱う. しかし, 通常, 文書内には複数意味を持つ単語（多義語）が存在し, 複数の単語が同じ意味を持つ場合（同義語）もある. これらの語を考慮することで, 精度を向上させるのがこの手法である.

本稿では, 共起語を考慮して EM アルゴリズムでの文書分類の精度を向上させる. EM アルゴリズムを用いた文書分類方法としては Nigam らの研究がある [5]. 単純ベイズ法に EM アルゴリズムを組み合わせた文書分類を提案しており，旋律の分類・検索 [10] やタイムスタンプの推定 [8] にはきわめて有用であることが知られる．しかし，単純に正規分布に従う確率密度を組み合わせて利用する方法ではきわめて低い分類精度しか達成できず，さまざまに変形を加えた試みを提案している．この変形はテキストデータに生じる特性に基づくものではなく，むしろ確率分布の特性をどう活かすかという視点に立つものが多い.

EM アルゴリズムを単純に適用すると, 精度が逆に悪化することが知られている [11]. つまりある地点まで精度が向上しても, 最終的にはそれよりも低い精度で収束する. 場合によってはラベル付き文書のみから生成した分類規則の制度よりも低い精度に収束することもあるため, 本稿では繰り返し回数を 20 回までとした.

2.3 EM アルゴリズムを用いたベイズ分類

本稿では, トピックをラベルとする訓練データを用いて, 未分類テキストデータを分類するための EM アルゴリズムを示す. 本稿では,EM アルゴリズムと単純ベイズ法を組み合わせた手法を用いる.

2.3.1 単純ベイズ法による文書分類

文書 d をトピック集合 $\mathcal{C} = \{ C_1, \ldots, C_n \}$ に分類する. ベイズ規則による分類とは, 文書 d がトピック C に属する確率 $P(C|d)$ の確率分布を求めることである. 排他的な分類の場合, 最大事後確率をとるトピック C へ文書 d を分類することで分類のミスを抑えることができる. 以下では, トピックラベルを次のように定める.

$$C_k = argmax_{C \in \mathcal{C}} P(C|d)$$

ベイズ規則を次のように定義する.

$$P(C_k|d) = P(C_k) \times \frac{P(d|C_k)}{P(d)} \tag{2.1}$$

すなわち, ベイズ規則での分類規則生成（訓練）は訓練データ集合から, 確率分布 $P(C_k), P(d), P(d|C$
を推定することである.

しかし, 文書ベクトル d=$(w_1, \ldots, w_m)$ はほぼすべての文書において異なり, $P(d|C_k)$ や $P(d)$ の推定が問題であるため, 一般に, 特徴量 w_j の出現は, 統計的に他の単語出現とは独立であるという仮定をおき, 各文書を単語の集合と考える単純ベイズ法を使うことが多い. 単純ベイズ法では $P(d|C_k)$ を以下の形式に分解して考える.

$$P(d|C_k) = \prod_{i=1}^{|d|} P(w_i|C_k)$$

これにより, 文書主導の排他的分類の場合, ベイズ規則は以下のように表すことができる.

$$P(C_k|d) = P(C_k) \times \prod_{i=1}^{|d|} P(w_i|C_k) \tag{2.2}$$

また, ここでは文書内での単語の出現頻度は考慮せず, 単語の出現有無のみを考えるバイナリ独立モデルを用いる.

バイナリ独立モデルとは, 単語の出現頻度を 0 か 1 で表現するもので, 文書内に出現したときは 1, 出現しなかったときは 0 とするものである.

例題 1 ベイズ手法によるクラス分類例を示す. トピック C を,"$C = \{C_1, C_2\}$" とし, それぞれが文書 d_1,d_2 として与えられているとする.（$D = \{d_1, d_2\}$）
各トピックの語集合を,

$C_1 = \{a, b, c\}$ $C_2 = \{a, d, e\}$

として, 未分類文書（d_3）を分類する.

単純ベイズ法によりトピックへ分類を行う. 定義より,$P(C_1|d_1) = P(C_2|d_2) = 1$, $P(C_1|d_2) = P(C_2|d_1) = 0$, $P(C_1|d_3) = P(C_2|d_3) = 0$ である. ベイズ規則を用いて,

$$P(C_k|d_3) = P(C_k) \times \frac{P(d_3|C_k)}{P(d)}$$

を最大化するトピック C_k を求める. $D = \{d_1, d_2, d_3\}$ には 3 件の文書が含まれおり, トピック C_1,C_2 にはそれぞれ 1 件ずつ含まれている. したがって,

$$P(C_k) = \frac{1 + \sum_{j=1}^{|D|} P(C_k|d_j)}{|C| + |D|}$$

$$P(C_1) = P(C_2) = \frac{1+1}{2+3} = \frac{2}{5}$$

文書	語群	単語数
d_1	a, b, c	3
d_2	a, d, e	3
d_3	a, b, c, f	4

トピック	$P(a\|*)$	$P(b\|*)$	$P(c\|*)$	$P(d\|*)$	$P(e\|*)$	$P(f\|*)$
C_1	2/4	2/4	2/4	1/4	1/4	1/4
C_2	2/4	1/4	1/4	2/4	2/4	1/4

各確率は, スムージングを行っている. この例題では, 確率が 0 となるのを防ぐため, 単純に分母と分子に 1 を足すことにする.

単純ベイズ法の仮定から,

$$P(C_k|d_3) = P(a|C_k) \times P(b|C_k) \times P(c|C_k) \times P(d|C_k) \tag{2.3}$$

ここで $P(C_1|d_3), P(C_2|d_3)$ をそれぞれ求める.

$$P(C_1) \times P(d_3|C_1)$$
$$= P(C_1) \times P(a|C_1) \times P(b|C_1) \times P(c|C_1) \times P(f|C_1)$$
$$= \frac{2}{5} \times \frac{2}{4} \times \frac{2}{4} \times \frac{2}{4} \times \frac{1}{4} = 0.0125$$
$$P(C_2) \times P(d_3|C_2)$$
$$= P(C_2) \times P(a|C_2) \times P(b|C_2) \times P(c|C_2) \times P(f|C_2)$$
$$= \frac{2}{5} \times \frac{2}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} = 0.0031$$

これより,$P(C_k|d_3)$ を最大化する C_k は C_1 となり, 文書 d_3 はトピック"C_1"に割り当てられた.

また,d_3 のトピック所属確率 $P(C_k|d_3)$ は,

$$P(C_1|d_3) = P(C_1) \times P(d_3|C_1)/P(d)$$
$$= \frac{0.0125}{0.0125 + 0.0031} = 0.80$$
$$P(C_2|d_3) = P(C_2) \times P(d_3|C_2)/P(d)$$
$$= \frac{0.0031}{0.0125 + 0.0031} = 0.20$$

2.3.2 EM アルゴリズム

EM アルゴリズムとは, データの欠損部分を最尤推定により求め, 欠損部分が分かればその形は単純かつ解析的に表現できるという仮定を置く. Expectation（期待値）,Maximization（最大化）は, それぞれ欠損値の推定, 期待値を得る過程を与えるパラメータの推定に対応している. この E ステップと M ステップを繰り返すことにより, モデルの対数尤度を最大化するパラメータを求める手法である.

EM アルゴリズムを文書分類に適用するために次を用いる.

1. 入力：ラベル付文書, ラベルなし文書
2. ラベル付文書のみから単純ベイズ分類規則 $\hat{\theta}$ を生成する
3. 以下のステップを一定回数, または分類規則が収束するまで繰り返す
   (a: E-step) 現在の分類規則 $\hat{\theta}$ を使用し，ラベルなし文書を各トピックへ分類する ($P(c_j|d_i;\hat{\theta})$)
   (b: M-step) 推定された事後確率（分類結果）による最尤推定を利用して, 分類規則 $\hat{\theta} = P(D|\theta)P(\theta)$ を再度生成する.
4. 出力：分類規則 $\hat{\theta}$

本稿では $P(w_i|C_k)$（分類規則）を以下の式で求める.

$$P(w_i|C_k) = \frac{1 + \sum_{j=1}^{|D|} N(w_i, d_j)P(C_k|d_j)}{|V| + \sum_{i=1}^{|V|}\sum_{j=1}^{|D|} N(w_i, d_j)P(C_k|d_j)} \tag{2.4}$$

ここで D は文書データ全体を表し、w_i はデータ内の各単語を表す. また $N(w_i, d_j)$ は文書 d_j における単語 w_i の発生回数であるが, 本稿では出現の有無により 0 か 1 の値をとる. さらに,$P(C_k|d_j)$ は前述の文書 d_j がトピック C_k に属する確率であり, ラベル付けされたデータに関しては, そのラベル付けられたトピック C_m においては,$P(C_m|d_j) = 1$ であり, それ以外のトピックに対しては 0 をとる. ラベルなしデータに関しては, 最初は全カテゴリに対して 0 であるが, 最初は通常のベイズ分類により, その後は EM アルゴリズムの E-step により, 徐々に適切な値へと更新される. 式 (2) と (3) により EM アルゴリズム内で分類規則を生成する. 同様に $P(C_k)$ は以下のように与えられる.

$$P(C_k) = \frac{1 + \sum_{j=1}^{|D|} P(C_k|d_j)}{|C| + |D|} \tag{2.5}$$

式 (3),(4) は, それぞれ $P(w_i|C_k), P(C_k)$ のスムージングを行っている.

例題 2 例 1 でのトピック推定の結果,

$$P(C_1) = \frac{1+2}{2+3} = \frac{3}{5}$$

トピック	$P(a\|*)$	$P(b\|*)$	$P(c\|*)$	$P(d\|*)$	$P(e\|*)$	$P(f\|*)$
C1	3/8	3/8	3/8	1/8	1/8	2/8
C2	2/4	1/4	1/4	2/4	2/4	1/4

$$P(C_2) = \frac{1+1}{2+3} = \frac{2}{5}$$

に変わる. 条件確率にも変化が起こる. ここで,

$$P(C_k|d_3) = P(C_k) \times \frac{P(d_3|C_k)}{P(d)}$$

を最大化する C_k を求めるため $P(C_k|d_3)$ を計算する.

$$P(C_1) \times P(d_3|C_1)$$
$$= P(C_1) \times P(a|C_1) \times P(b|C_1) \times P(c|C_1) \times P(f|C_1)$$
$$= \frac{3}{5} \times \frac{3}{8} \times \frac{3}{8} \times \frac{3}{8} \times \frac{2}{8}$$
$$= 0.0079$$

$$P(C_2) \times P(d_3|C_2)$$
$$= P(C_2) \times P(a|C_2) \times P(b|C_2) \times P(c|C_2) \times P(f|C_2)$$
$$= \frac{2}{5} \times \frac{2}{4} \times \frac{1}{4} \times \frac{1}{4} \times \frac{1}{4}$$
$$= 0.0031$$

また,d_3 のトピック所属確率 $P(C_k|d_3)$ は,

$$P(C_1|d_3) = P(C_1) \times P(d_3|C_1)/P(d)$$
$$= \frac{0.0079}{0.0079 + 0.0031} = 0.72$$
$$P(C_2|d_3) = P(C_2) \times P(d_3|C_2)/P(d)$$
$$= \frac{0.0031}{0.0079 + 0.0031} = 0.28$$

この結果、$P(C_k|d_3)$ を最大化するトピック C_k は C_1 であり,"d_3"は再度トピック C_1 に割り当てられる. EM アルゴリズム部で, 変動が起きなくなるまで計算を繰り返す.

2.4 共起語を考慮した EM アルゴリズム

2.4.1 共起語

本稿では, 語の対 (w_i, w_j) の d における共起度 $co(w_i, w_j)$ を次のように定義する[1] : $co(w_i, w_j) = \sum_{s \in d} |w_i|_s |w_j|_s$

ここで $|x|_s$ は文 s における要素 x の出現回数を表し, x が語の場合には $|x|_s$ は文 s 中の語 x の出現回数を表す. 式 (5) は, ある文 s に出現した語 w_i は s 中のすべての語 w_j と共起しているとみなした共起頻度を表す. すでに述べたように，本稿では語の共起を文単位で考慮し，複数の文にわたる影響を考えない.

2.4.2 共起語を含む EM アルゴリズム

本章では,EM アルゴリズムを拡張し，単語同士の共起関係を考慮する. 基本的なアイデアは単純である. すなわち，未分類文書 d が, トピック C_k に割り当てられたとき, トピック C_k に属する単語 $w_1, ..., w_n$ のいずれかと, 文書 d 内において共起し, しきい値以上の共起回数を持つ語 x をトピック C_k に追加する. この EM アルゴリズムを共起語で拡張する手法が，実際に分類精度を向上させるものとなるであろうか？一般にテキストデータでは，語の同義性や類義性からいくつかの語が同時に生じる可能性が高い. つまり共起語は，特に高頻度の語の場合は強い相関性があると考えることができ，実際これを手がかりに文書分類の研究がなされている [3, 6].

更に，本稿では追加する共起語の頻度をある一定以上の割合とする. しきい値 5 割で, 文書内の最大共起頻度が 10 回ならば, 追加語は共起頻度が 5 回以上のものとなる. これは, 追加語の共起頻度に制限を設けることによって, 追加語とトピックの語集合との相関の強さを操作するためである. しきい値を高く設定すれば, 相関的な語が追加されるので, より精度が向上するはずである. 逆に, しきい値を低くすれば, トピックと相関の低い語が追加される機会が多くなるので, 高く設定した場合よりも精度は悪化すると考える.

例題 3 共起語を考慮に入れた EM アルゴリズムによるクラス分類例を示す.
ここでは, 例 1 と同じデータを用いて例を示す. トピック C を,"$C = \{C_1, C_2\}$" とし, それぞれが文書 d_1,d_2 として与えられているとする. ($D = \{d_1, d_2\}$)
各トピックの語集合を,

$C_1 = \{a, b, c\}$ $C_2 = \{a, d, e\}$

として, 未分類文書（d_3）を分類する.

[1]本稿では共起を文単位で考えている. そのため, 共起を 3 単語まで広げる必要がない. つまり,3 単語で共起していれば, 必ずそのうちの 2 単語も共起していることになる. しかし, これは同時に, 熟語を考慮しないということにもなる. KeyGraph[6] では文章に限っていないために"共起語集合"を扱う.

はじめに単純ベイズ法によりトピックへ分類を行う. 計算方法および結果は, 例 1 と同様である.

次に推定されたトピックに共起語を追加する. 追加する語は, トピックの語集合の単語のいずれかと, 文書 d_3 内において共起している語が対象となる. このとき, 対象となる語がすでにトピックの語集合に含まれている場合は, 追加を行わない.

ここで, 文書 $d_3 = \{a, b, c, f\}$ のとき, 語対 (w_i, w_j) の共起度 $co(w_i, w_j)$ はそれぞれ,

$$co(a, f) = 3$$

$$co(a, b) = co(b, c) = co(b, f) = 2$$

$$co(a, c) = co(c, f) = 1$$

である. しきい値を 80%に設定すると, 追加対象となる共起語は, 追加語の共起度が最大頻度 × しきい値 $= 3 \times 0.8 = 2.4$ よりも大きいため, 共起度 3 の語のみが対象となる. これより, 追加する語の条件は, トピック $C_1 = \{a, b, c\}$ のいずれかと共起する語, かつ共起度 3 の語となる. d_3 内において, この条件を満たす語は"f"のみであり, これをトピック C_1 の語集合に追加し,$C_1 = \{a, b, c, f\}$ とする. 以降, トピック C_1 の語集合を $C_1 = \{a, b, c, f\}$ として分類を行う.

新たに定まった確率値を用いて, 再度 d_3 の所属確率を求める (EM アルゴリズム).

さらに, 推定されたトピックに共起語を追加する. 前述と同様に共起語の追加作業を行うが, すでに対象となる語が含まれている場合は追加作業はない.

最後に，EM アルゴリズム部で再度計算を繰返し, 変動が起きなくなるまで計算を繰り返す.

2.5 実験

2.5.1 実験データ

本稿では，シェークスピアによる戯曲 12 作品をテストコーパスとして使用し, タイトルをトピックとして考える. 各タイトルはいずれも 5 章 (chapter) から構成され各章は場 (Scene) からなる. 実験で用いたタイトルは次の 12 作品である.

1. 真夏の夜の夢 (全 5 章 9 場)
2. 終わりよければすべてよし (全 5 章 23 場)
3. お気に召すまま (全 5 章 22 場)
4. シンベリン (全 5 章 26 場)
5. 恋の骨折り損 (全 5 章 10 場)
6. から騒ぎ (全 5 章 18 場)

7. ペリクリーズ (全 5 章 20 場)

8. 間違いの喜劇 (全 5 章 11 場)

9. ヴェニスの商人 (全 5 章 19 場)

10. ウィンザーの陽気な女房たち (全 5 章 23 場)

11. じゃじゃ馬ならし (全 5 章 14 場)

12. テンペスト (全 5 章 9 場)

初期訓練データを各トピックの第 1 章のすべての場から生成し, 2 章以降の全 168 場をテストデータとして使用する (表 1 参照). それぞれの場のタイトルを隠して分類を行い, 後に適切に分類がされたかを判断する. 各場データはあらかじめステミング [7] および不要語処理を行う. また, 初期訓練データに Zipf の法則を適用し, 上位 30 単語をトピック単語とする. 各トピックの構成と, ストップワードを除去した後の各場の単語数を表 1 に示す.

共起語の計算は, 各場ごとに行う. 本稿では, 共起語を各場内において, 出現頻度が 2%以上の語のみを計算する. これは, 相関性の極めて低い単語を, 共起語として算出しないためである.

2.5.2 実験手順

提案手法の評価を以下 3 点において行う.

- 精度への影響
- しきい値による正答率の変化
- 多数トピックへの共起語の削除の有効性

提案手法の有効性を示すため，本実験では，EM アルゴリズムによる分類と, 提案手法である共起語を考慮に入れた EM アルゴリズムの正答率の変化を比較することで. 提案手法を検証する. 正答率は次式で定義される.

$$\frac{\text{正しく分類された総文書数}}{\text{総文書数}} \times 100\ (\%) \tag{2.6}$$

推定されたトピックと実際のトピックとの比較を行い, 精度への影響を検証する. しきい値を 10%から 100%と変化させて場合と, しきい値を設けなかった場合についても実験を行う. 最後に, 多数トピックへの共起語の削除の有効性を, 同じしきい値で, 削除した場合と, 削除しなかった場合で比較し検証する.

また, 本稿では EM アルゴリズムの繰り返し回数を, 0 回（単純ベイズ部）,5 回,10 回,15 回,20 回の 5 パターンについて示す.

表 2.1: データの構成とサイズ

トピック	場	1 章	2 章	3 章	4 章	5 章
1	1	512	146	120	123	219
	2	198	85	242	24	-
2	1	462	123	7	39	16
	2	224	26	69	43	31
	3	482	154	8	167	184
	4	-	21	25	14	-
	5	-	36	49	52	-
	6	-	-	44	-	-
	7	-	-	15	-	-
3	1	322	30	12	89	36
	2	429	35	189	12	49
	3	280	8	46	83	22
	4	-	13	26	-	127
	5	-	111	65	-	-
	6	-	34	-	-	-
	7	-	55	-	-	-
4	1	416	27	37	8	11
	2	78	118	39	240	14
	3	114	68	58	28	52
	4	350	7	115	22	97
	5	227	-	90	-	296
	6	479	-	50	-	-
	7	-	-	9	-	-
5	1	576	137	103	97	76
	2	311	-	140	96	516
	3	-	-	-	243	-
6	1	518	185	122	156	186
	2	69	23	67	47	48
	3	158	126	46	-	18
	4	-	-	86	-	77
	5	-	-	47	-	-
	6	-	-	25	-	-
7	1	110	90	60	53	140
	2	399	31	21	68	58
	3	302	25	23	54	-
	4	92	58	-	3	-
	5	353	122	-	109	-
8	1	376	66	76	71	238
	2	233	108	99	47	-
	3	-	-	-	46	-
	4	-	-	-	109	-
9	1	624	17	68	249	165
	2	346	106	170	13	-
	3	-	11	19	-	-
	4	-	26	39	-	-
	5	-	30	51	-	-
	6	-	42	-	-	-
	7	-	34	-	-	-
	8	-	25	-	-	-
	9	-	52	-	-	-
10	1	444	115	60	40	20
	2	40	148	42	118	7
	3	258	47	115	9	10
	4	304	-	50	47	1
	5	-	-	60	81	137
	6	-	-	-	30	-
11	1	336	128	53	117	84
	2	341	139	131	59	112
	3	-	224	-	103	-
	4	-	-	-	54	-
	5	-	-	-	43	-
12	1	187	231	49	146	200
	2	970	108	87	-	-
	3	-	-	66	-	-

2.5.3 実験結果

表 2 に, ベイズ手法に EM アルゴリズムを組み合わせた分類の正答率と, 共起語を考慮に入れた EM アルゴリズムの各しきい値に対する正答率を示す. 表 3 に,EM アルゴリズムと, 共起語を考慮に入れた EM アルゴリズムの各実験結果のうち, 最も精度の高かったしきい値との分類結果の違いを示す.

また, 共起語の追加に対する精度の変化過程を評価するため, 代表的な例について考察する. 表 4 に, トピック単位での正答率を示す. 削除を繰り返し, 最終的に残った追加語を表 5 に示し, 削除が行われた語とそのタイミングを表 6 に示す. 分類の変化の例を表 6 に示す. この表 7 について, ○は EM アルゴリズムでの不正解が, 提案手法に置いて正答に変わったデータ. ×は逆に不正解に変わったデータを示す. また, 表 8 にはしきい値による追加語が生じた件数を示し, 表 9 に多数トピックへの追加語の削除についての正答率を示す.

表 2.2: 正答率

%	Bayes	EM(5)	EM(10)	EM(15)	EM(20)
Bayes + EM	29.17	27.38	28.57	38.10	38.10

%	しきい値	EM(0)	EM(5)	EM(10)	EM(15)	EM(20)
B	100	38.10	30.36	29.76	48.81	48.81
a	90	38.69	27.38	29.76	50.00	50.00
y	80	41.07	29.76	30.95	54.76	54.76
e	70	37.50	26.79	30.95	51.19	51.19
s	60	32.74	21.43	28.57	47.62	47.62
+	50	39.88	27.38	30.36	45.24	45.24
共	40	27.38	25.60	26.79	43.45	43.45
起	30	28.57	24.40	21.43	36.31	36.31
+	20	32.14	22.02	22.62	38.10	38.10
E	10	32.14	23.81	21.43	36.90	36.90
M	0	31.55	23.21	21.43	35.71	35.71

表 2.3: EM アルゴリズムと共起語を考慮に入れた EM アルゴリズム（しきい値 80%）の正答の違い

		Bayes+EM		繰返し
B	個数	正解	間違い	
a	正解	39	30	0
y	間違い	10	89	
e	正解	29	21	5
s	間違い	6	112	
+	正解	17	21	10
共	間違い	5	125	
起	正解	59	33	15
+	間違い	5	71	
E	正解	59	33	20
M	間違い	5	71	

表 2.4: トピックごとの正答率

しきい値 80 %						
トピック	EM(0)	EM(5)	EM(10)	EM(15)	EM(20)	総場数
1	57.14	0.00	14.29	71.43	71.43	7
2	20	20	25	40	40	20
3	42.11	36.84	36.84	47.37	47.37	19
4	35.00	35.00	35.00	55.00	55.00	20
5	75.00	12.50	25.00	62.50	62.50	8
6	40.00	40.00	40.00	80.00	80.00	15
7	33.33	26.67	26.67	46.67	46.67	15
8	33.33	33.33	44.44	66.67	66.67	9
9	47.06	29.41	29.41	47.06	47.06	17
10	63.16	47.37	47.37	63.16	63.16	19
11	25.00	16.67	0.00	25.00	25.00	12
12	42.86	28.57	28.57	85.71	85.71	7

Bayes+EM						
トピック	EM(0)	EM(5)	EM(10)	EM(15)	EM(20)	総場数
1	42.86	42.86	42.86	71.43	71.43	7
2	20.00	10.00	15.00	25.00	25.00	20
3	21.05	26.32	26.32	26.32	26.32	19
4	25.00	25.00	20.00	35.00	35.00	20
5	25.00	12.50	12.50	25.00	25.00	8
6	33.33	20.00	33.33	53.33	53.33	15
7	13.33	20.00	20.00	26.67	26.67	15
8	33.33	22.22	33.33	44.44	44.44	9
9	35.29	23.53	23.53	29.41	29.41	17
10	52.63	57.89	63.16	68.42	68.42	19
11	8.33	16.67	8.33	8.33	8.33	12
12	57.14	71.43	57.14	71.43	71.43	7

表 2.5: 共起語を考慮した EM アルゴリズム しきい値 80% 追加語

文書	分類先	追加語	繰返し		文書	分類先	追加語	繰返し
1	7	fair	2		71	5	beatric	3
1	3	king	4		72	5	hath	0
2	12	fair	8		72	5	fanci	0
3	3	hear	3		72	10	fanci	2
3	3	honour	3		74	9	light	0
3	6	hear	4		74	6	light	2
29	11	our	2		81	5	ill	2
29	10	our	3		81	9	ill	10
29	8	make	4		84	9	twenti	0
39	11	ay	3		87	10	ey	0
39	11	young	3		87	10	clock	0
39	11	man	3		88	10	sir	4
39	11	mine	3		133	4	bianca	3
39	11	sir	3		139	4	presum	0
39	4	find	8		143	7	blame	2
40	8	call	3		143	3	blame	3
41	10	thee	0		153	5	thee	4
44	11	morrow	4		161	10	brook	0
51	3	great	0		163	10	betrai	0
52	5	write	0		163	10	amaz	0
60	8	man	5		165	11	ann	4
71	9	beatric	2					

表 2.6: 削除語

文書	分類先	追加語	繰返し	文書	分類先	追加語	繰返し
12	4	love	11	127	2	hugh	5
23	9	master	4	127	8	husband	11
31	2	hath	11	149	3	art	4
39	11	find	3	149	3	posthumu	4
39	11	good	3	153	6	thy	0
40	8	give	3	156	6	live	3
40	5	desir	4	157	9	thou	2
43	6	fit	5	158	11	brought	3
44	8	morrow	3	160	7	night	4
103	10	wit	5	165	2	mistress	5
124	10	half	4				

2.5.4 考察

表 2 より明らかなように, 語の共起性を EM アルゴリズムに融合させてテキスト分類を行うことは, 精度を向上させることができる. また, 本稿では, 追加語を一定回数以上の共起頻度を有する語に限定し, より効果的なつながりを重視した. しきい値を 80%にした場合に最も良い精度を得たが, 100%,90%,70%の場合においても,EM(15),EM(20) において,10%以上の精度向上が見られる. 逆に, このしきい値を設けなかった場合,10%と低く設定した場合において, 精度の低下が見られる. 共起を考えるにあたり, 相関の高い語を追加することが有効であることがいえる.

本稿で提案したように, EM アルゴリズムに共起語を考慮に入れて分類を行うことが, 有効的であったことが確認できる.

共起語の影響

本実験において, 誤分類による共起語の追加が正答分類による追加の数を上回ったが, 精度が向上した. 誤分類の追加語の影響より, 正答分類での追加語の影響が強いことが挙げられる. これは, トピックの語集合は, トピックへの依存が比較的強い語で構成されていることからくる. 文書の構成は, 単語の意味を考慮して書かれている場合が多く, 本来のトピックの語集合との共起と, 誤分類先の語集合の共起は違う意味を持つ. そのため, 誤分類の追加語において, トピック特有の単語が追加されるケースは少なく, 誤分類においては汎用性のある単語が追加されるケースが多い. また, このような追加語は多数トピックの共起となるケースも多く, 本稿で用いた手法における, 削除の対象となる.

しかし, 誤分類は必ずしも精度の悪化を招く要因とはならない. 共起が起こるということは相関を持つ語であるということである. 本稿では, 第 1 章の単語に Zipf の法則を適用し, その頻出上位 30 単語をトピックの語集合とした. その際,30 単語には入らなかったが, トピックの単語である語も存在する. 表 5 より, 文書（60）において, 単語（man）が追加語としてトピック 8 に追加されている. この文書（60）は本来はトピック 3 に分類される文書である. しかし, この単語（man）が文書（145）の中で存在しており,

EM アルゴリズムで不正解だった分類が, 本稿の提案手法において正答に移り変わった. このように, 誤分類による追加は必ずしも精度の悪化を招く要因とはならない.

また, 共起語の追加で精度が悪化するケースであるが, 文書（97）において著しく悪化しているケースがある. この文書（97）は本来はトピック 1 に分類される文書であるが, 提案手法において, トピック 5 に分類されることで精度が悪化している. この文書はトピック 1 の 4 章:場 2 に当たる文書であり, 総単語数は 24 である. このような少ない単語数からなる文書は,1 単語のトピック推定に与える影響は他のものに比べて大きくなるのだが, 文書（72）において, この文書（97）にも含まれる単語（hath）がトピック 5 に分類されている.

表 4 より, 各トピックごとについても,EM(0),EM(15),EM(20) でほぼ全てのトピックで正答率の向上が見られる. これは, 直接共起語の追加がおきていないトピックに対しても向上しており, 共起語の追加は, 追加の起きたトピックのみならず, その他のトピックの精度にも影響を与えると見れる. 他トピックへの共起語の追加により, 各トピックに対する所属確率にも変動が起き, 精度にも影響が出る. 本提案手法である共起語の考慮は, 特定トピックのみでなく, トピック全体での精度向上に役立つといえる.

また, 共起語を考慮に入れた EM アルゴリズムにおいて,EM(15),EM(20) で最も良い精度となっている. 繰り返し回数による精度向上の幅に, 途中での追加語が影響していることは考えられるが, EM アルゴリズムの結果においても同様の変化がおきている. そのため,EM(15),EM(20) で最も良い精度となる大きな要因としては, EM アルゴリズムの繰り返し計算により, 正しいトピックに分類されていくことが大きいと考えられる.

これら, 正答分類による追加語と, 誤分類における追加語の影響で, 本提案手法において精度の向上が見られた.

しきい値の影響

本稿では, 実験を行う際に, 追加する共起語の共起頻度にしきい値を設けた. 表 2 より, しきい値を高くし, 相関の強い語を追加することの有効性が確認できる. このときの追加語のしきい値としては, 本実験において,80%の精度が最も良い. これは,80%に設定する事で,100%の場合よりも相関の強い語が多く追加される理由による. また, しきい値を下げると, 依存性の弱い単語, つまりどの文書でも出現してくるような単語が多く追加される. 意味のない単語の追加, 各トピックの差が弱まる事で精度が悪化し, また誤分類の追加語が増える事となる. 追加件数にあまり差がないのは, 同一文書において, 分類先が変わるたびに同じ単語の追加が起きているためである.

多数トピックへの共起語の追加

本稿では, 多数トピックへの共起語の追加を制限し,3 つ以上のトピックへの追加が起こるケースにおいて, その追加語を追加した全てのトピックから削除している. 表 9 より, 本実験において, その手法を用いたことが有効であったことが分かる.

表 2.7: EM アルゴリズムと共起語を考慮に入れた EM アルゴリズム の分類の違い

文書	繰り返し回数				
	0	5	10	15	20
144	○	-	-	-	-
145	-	-	-	○	○

文書	繰り返し回数				
	0	5	10	15	20
96	-	×	-	-	-
97	-	×	×	×	×

表 2.8: 共起語の追加作業が起きた件数

しきい値	追加総数
100%	79
80%	108

表 2.9: 多数トピックへの共起語の削除評価

Bayes+共起+EM　しきい値= 50% 削除なし					
	EM(0)	EM(5)	EM(10)	EM(15)	EM(20)
正答数	54	46	49	74	74
正答率	32.14	27.38	29.17	44.05	44.05

Bayes+共起+EM しきい値= 50% 削除あり					
	EM(0)	EM(5)	EM(10)	EM(15)	EM(20)
正答数	67	46	51	76	76
正答率	39.88	27.38	30.36	45.24	45.24

2.6 結び

本稿では, 共起語を考慮に入れた EM アルゴリズムによるテキスト分類の手法を提案した。テストコーパスを用いた実験によって, EM アルゴリズムに対して, 最も良い精度を得た結果で, EM(0) で 11.9%, EM(5):2.38%, EM(10):2.38%, EM(15):16.66%, EM(20)16.66%と精度が向上したことを確認した. また, 追加語のしきい値を高くし, より効果的なつながりを重視しすることで, しきい値を設けなかった場合, 低く設定した場合よりも高精度の結果を得た. 共起を考えるにあたり, 相関の高い語を追加することが有効であり, 本稿によって, 相関的な共起語を考慮に入れることにより, テキスト分類の精度を向上させることができることを確認した.

第3章　単語分布からのトピック推定

3.1　前書き

過去一世紀以上にわたる論争の一つに著者推定問題がある。何らかの特徴を捉えて著者の同定・区別を行おうとするもので, シェークスピア実在論争やグリコ森永事件の脅迫状分析等がその応用例である [14]．著者推定・分析を行うためには, 文体の計量的特長 (stylometry), 例えば語長・文長・語数や機能語 (while, on などの不要語記号) などを調べる方法があるが, 同一筆者でも差が大きく特徴が有効とはいいがたい [16].

これと並んで興味あるものがトピック推定問題である．トピック (topic) とは興味ある事柄や出来事を言い, テキスト文書を解析し何のトピックを論じるものかを推定することをトピック推定問題という．この技術は，文書の自動格納・自動分類や自動要約に主要な手がかりを与え, また背景や領域の推定による文脈情報を付加することで，情報検索の効率向上に大きなヒントを与えることができる．

これまでの研究結果から，テキストから直接有用な情報を抽出する方法では, 著者固有の性質よりもトピックとの関連性を論じるほうが分析しやすいことが知られている [16]．著者トピック語モデル (Author Topic Model) とは著者推定がトピック（テーマ）の選定に確率分布に従うことをいう．これに対して，同一著者の下では, 各トピックは対応する語集合の多項式分布確率で表わされるとする**トピック語モデル** (Topic Word Model) が議論されることが多い [20]．従って，トピック語モデルが正しければ, 語の分布を調べることでトピック推定が可能であり, 具体的な推定手順を与える論拠となる．一般に, 文書は複数トピックを含むが, 本稿では 文書とトピックを同一視し, トピック推定を効率よく実現する手法を検討する．

これまで情報検索では，文書中の語をベクトルで表わし，ベクトル類似の問題に帰着するベクトル空間モデルが知られる [13]．モデル化が単純であり類似度も簡単に算出できることから, 広く利用されているが, 解が重み付け方法に依存し, また数万に至る高次元処理が必要であることから，性能および精度に差が生じる．このため精度を維持したままで効率向上を目的とした次元縮小技法が知られている [13].

本稿では，同一著者の下でトピック語モデルの検証を行い，次に次元縮小技法をトピック語モデルに適用し，トピック推定に有用であることを論じる．

第 2 章ではトピック語モデルと評価方法を述べる．第 3 章では次元縮小手法を要約し, 適用手法を論じ，さらに第 4 章で実験によりその有用性を示す.

3.2 トピック語モデル

トピック語モデルとは, 同一著者の作品（トピック）には特有の語分布が対応し，各語は多項式分布に従って確率的に選ばれるという特徴を仮定することをいう．これが正しければ，一般には, トピック上の語分布を比較することでトピック推定が可能になる．

テキスト情報は語の並びとして構成されるが, 語 (word) をどのように設定するかは自明ではない．英語では (空白などの) 特殊文字で区切られた文字列を単語と呼ぶが, 複合語 (New York 等のように複数の単語からなる語) や共起性の強い語 (連語や慣用句) 等を考慮するかどうかは, 分析結果に大きな影響を与える．n グラム (n-gram) モデルでは, 連続する n 単語をまとめて語とみなすが, 単語の区切りを無視して数え上げるため, 多くのミスを含む可能性がある．反面, 複合語や共起性問題を取り扱うことができる．日本語では形態素を基本とする．形態素 (Morphology) とは，これ以上に細かくすると意味を失う最小の文字列を言う．文章を形態素に分解する処理を形態素解析と呼ぶ．形態素は単語に対応するが, 英語と同様に複合語・共起語の対応を考える必要がある．

本稿では，トピック語モデルを検証するため, 英文テキストに対して n グラムモデル ($n = 1, 2$) を用いて語分布を調べる．テキスト文書に出現する単語を，予めステミングおよび不要語処理を行い自明情報を取り除いたあと, トピックの一部から抽出した教師データと, 残りから抽出したテストデータの語を調べその出現頻度分布を比較する．

評価を行う際にカイ 2 乗検定の X^2 値を用い，教師データとテストデータの分布を比較し独立性を調べる．このため教師データの語 w_i の頻度を期待値 a_i，テストデータの語 w_i 頻度を観測値 b_i として，以下の式を用いて評価する．

$$X^2 \quad = \quad \sum_{i=1}^{n} \frac{(b_i - a_i)^2}{|a_i|} \tag{3.1}$$

上式では，X^2 値が小さいほど分布が依存することをあらわしている．本稿ではデータ量の不均等を考慮し X^2 値の下位 3 つを本実験での正答とする．ここで正答率は，検定文書中の適合する数を p，文書総数 (総場面数) を q とするとき次式で定義される．

$$\frac{p}{q} \tag{3.2}$$

3.3 情報検索とデータ次元縮小

文書に含まれるテキスト情報を検索するには，出現する各語の (出現頻度等) 特徴を値としてベクトル化するベクトル空間モデルが一般的である [13]．一般にテキスト文書 d は, 出現する語 $w_1, .., w_n$ のベクトルで表現する:

$$d = (v_1, ..., v_n)$$

ここで v_i は語 w_i に対応する数値であり一般に出現頻度 (Term Frequency) であることが多い．このとき 2 つの文書 d_1, d_2 の類似度は出現数の分布を用いて定義され, これは内積 (d_1, d_2) によって算出できる．

この方法は，モデル化が単純であり類似度も簡単に算出できることから, 広く利用されているが, 解が重み付け方法に依存し, 次元数が数万にも及ぶ高次元データをそのまま扱うと，効率, 計算機容量の確保および即応性への対応が困難になる．このため, テキスト情報の次元を縮小し改善を図る次元縮小技法が知られている [13]．次元縮小技法では高次元文書ベクトルを低次元空間に射影し, この部分だけを検索対象とするため，効率よく探索範囲を絞り込むことができる．

次元縮小技法のうち, 潜在意味索引つけ (Latent Semantic Indexing) は，源データを用いて特徴値を算出するためきわめて高精度に縮小可能であるが，特徴値算出手続きの効率が悪くまた微小な変更でも再計算を要求することから動的な環境に利用できない．一方, ランダムプロジェクション (Random Projection, RP) 技法は，乱数技法により次元縮小するため, 次元縮小手続きの効率が良く，低次元空間に縮小するほど少なくて済む利点がある．またテキスト文書集合が増えても再計算を要求することがない．反面, 精度が悪く適用範囲に限界がある．

本稿では，RP 技法を用いて次元縮小を行う．以下では語数 d，文書数 N とし，$X \times N$ 語・文書行列 X を $k \times N$ $(k \ll d)$ の語・文書行列 X_{RP} に射影する．射影を行うため，$k \times d$ の RP 行列 $R = ((r_{ij}))$ を生成する．この際，行列 X の i 行 j 列の要素 X_{ij} は，文書 j における語 i の頻度を意味する．単語・文書行列 X の RP による次元縮小の計算は以下のように定義される[1]:

$$X^{RP}_{k\times N} = R_{k\times d} X_{d\times N} \tag{3.3}$$

また，第 i 行ごとに，RP 行列の要素 $r_{ij}, j = 1, ...$ は，発生確率 p に対して，次の分布に従うように決定する [12].

$$r_{ij} \;=\; \sqrt{3} \cdot \begin{cases} +1 & (p = 1/6) \\ 0 & (p = 2/3) \\ -1 & (p = 1/6) \end{cases} \tag{3.4}$$

この分布を取る行列の生成に対する計算量は $O(kd)$ であり，$(k \ll d)$ でもあることから，実際の処理時間は非常に少ない.

$d \times N$ 行列 X から任意の列ベクトルを取り出し，予め用意した訓練用データを用いて RP 行列を作成しテストデータを比較・評価する．

本研究では，RP を用いた検索では縮小率に伴う正答率の低下を評価する．

3.4 実験

3.4.1 準備

シェークスピアによる戯曲 10 作品 (英語テキスト) をテストコーパスとして使用し，作品をトピックと考える [15]．各タイトルはいずれも 5 章（chapter）から構成され，各章は場（scene）からなる．実験で用いる作品タイトルは次である．

[1]データの初期生成に対する計算量は $O(dkN)$ となる.

タイトル	構成
夏の夜の夢	全 5 章 9 場
お気に召すまま	全 5 章 22 場
シンベリン	全 5 章 26 場
ハムレット	全 5 章 20 場
オセロー	全 5 章 15 場
ジュリアス・シーザー	全 5 章 18 場
ジョン王	全 5 章 16 場
リチャード二世	全 5 章 19 場
ヘンリー八世	全 5 章 17 場
テンペスト	全 5 章 9 場

教師データとして各トピックの第 1 章の全ての場から単語を抽出し，また 2 章以降の全 138 場から単語を抽出してテストデータとして使用する．テストデータとしての場は，そのタイトルを隠して推定し，上位 3 トピックに正解が含まれているとき正しく推定されたと判断する．各トピックの構成と，不要語を除去した後の各場の単語数を表 2 に示す.

本実験では 1 グラム (実験 1), 2 グラム (実験 3) によるトピック語モデルの検証と，RP 技法による次元縮小の精度 (実験 2) を調べる．特に，次元縮小の精度調査 (実験 2) では縮小率に伴う正答率の低下を評価する．RP 技法では行列はランダムに生成されるため，10 回繰り返した正答を平均した値を本実験での正答率とする．

3.4.2 実験結果

表 3.1 に 実験 1 のおける推定の正答率を示し，表 3.2 に (実験 1 の) 各教師データの単語総数を示す．表 3.3 に各文書における本来の正答トピックと推定結果を○と×で示す．場の番号 1 から 33 までが 2 章，34 から 71 が 3 章，72 から 100 が 4 章，101 から 138 が 5 章の場を意味する．次元縮小による精度低下を評価するため，縮小した各次元における正答率と縮小前の正答率からの低下割合を表 3.4 に示す．同様に 2 グラム分布を用いる実験 3 におけるトピック教師データの総語数を表 3.6 に示し，その正答率を表 3.8 に示す.

表 3.1: 1 グラムモデルでの正答率

トピック	C1	C2	C3	C4	C5	
正答率	100	100	40	33.3	75	
トピック	C6	C7	C8	C9	C10	合計
正答率	86.7	100	46.7	76.9	100	72.46

表 3.2: 1 グラムモデルの教師データトピック総語数

トピック	C1	C2	C3	C4	C5
単語数	654	783	1171	1327	1132
トピック	C6	C7	C8	C9	C10
単語数	849	503	1120	1138	1067

3.4.3 考察（実験 1）

表 3.1 と表 3.2 より，教師データの大きさによる推定結果の違いはあるが，正答率が 72.46% であることから，本実験で用いるデータがトピック語モデルに従うといってよい．表 3.1 においてトピック C3，C4 の推定が低精度となっているが，表 3.7 の各トピック間の出現語の違いより，他トピックに対して偏った語の出現でないことが確認でき，特殊な分布ではないことが確認できる．この 2 つのトピックの推定が低精度となる要因として，(1) 第 1 章からの教師データ作成が困難であったこと，(2) 表 3.2 と表 3.7 より，他トピックに共通する語の出現が多いこと, (3) 語総数自体が多いことによる．このため，推定が語総数の低いトピックへと割り当てられ精度が低下する要因となる．

3.4.4 考察（実験 2）

表 3.4 により，140 次元への縮小（縮小率 98.59%）で精度低下が 19.30% 程度である．表 3.5 では, 判定が 10 回中過半数で変化が生じるとき「判定変化あり」とする[2]．表 3.5 より，変化なしの割合は高く，次元の縮小によるトピック語モデルの信頼性は変化がない．

3.4.5 考察（実験 3）

表 3.9 は，本実験で数多く推定された先のトピックとその推定数を示す．表 3.9 より，トピック C1，C6，C7 に推定されるテストデータが多い．しかし, 表 3.6 より，いずれも教師データの総語数が低いトピックである．また，表 3.10 は 1 グラムモデルと 2 グラムモデルの次元と，1 度だけ出現する語の頻度を示している．表 3.10 より，1 グラムモデルでは各トピックに有効で出現が考えられる語は 56.12%なのに対して，2 グラムモデルでは 8.47%である．他トピックに出現しない語は，そのままデータ量の違いの影響を受けやすく，偏りが生じやすい．従って 2 グラムモデル分布によって，トピック語モデルが検証できたとはいい難い．

[2]変化の度合いが 5 回ずつであれば判定不能とする．

表 3.3: 1 グラムモデルでの推定

場	正解トピック	推定不可	場	正解トピック	推定不可
1	1	○	70	10	○
2	1	○	71	10	○
3	2	○	72	1	○
4	2	○	73	1	○
5	2	○	74	2	○
6	2	○	75	2	○
7	2	○	76	2	○
8	2	○	77	3	×
9	2	○	78	3	○
10	3	○	79	3	×
11	3	○	80	3	×
12	3	○	81	4	×
13	3	×	82	4	×
14	4	×	83	4	×
15	4	○	84	4	×
16	5	○	85	4	×
17	5	×	86	4	×
18	5	○	87	4	×
19	6	○	88	5	○
20	6	○	89	5	○
21	6	○	90	5	○
22	6	○	91	6	×
23	7	○	92	6	○
24	8	○	93	6	○
25	8	○	94	7	○
26	8	○	95	7	○
27	8	×	96	7	○
28	9	○	97	8	○
29	9	○	98	9	○
30	9	○	99	9	○
31	9	○	100	10	○
32	10	○	101	1	○
33	10	○	102	2	○
34	1	○	103	2	○
35	1	○	104	2	○
36	2	○	105	2	○
37	2	○	106	3	×
38	2	○	107	3	×
39	2	○	108	3	×
40	2	○	109	3	○
41	3	×	110	3	○
42	3	×	111	4	○
43	3	×	112	4	○
44	3	○	113	5	○
45	3	○	114	5	○
46	3	×	115	6	○
47	3	×	116	6	×
48	4	×	117	6	○
49	4	○	118	6	○
50	4	×	119	6	○
51	4	○	120	7	○
52	5	×	121	7	○
53	5	×	122	7	○
54	5	○	123	7	○
55	5	○	124	7	○
56	6	○	125	7	○
57	6	○	126	7	○
58	6	○	127	8	×
59	7	○	128	8	○
60	7	○	129	8	○
61	7	○	130	8	×
62	7	○	131	8	×
63	8	×	132	8	×
64	8	×	133	9	○
65	8	○	134	9	×
66	8	×	135	9	○
67	9	○	136	9	×
68	9	○	137	9	×
69	10	○	138	10	○

表 3.4: 次元縮小における精度の変化

次元数	正答数	正答率	精度低下
9923	100.00	72.46	0.0 %
9000	98.10	71.09	1.89
5000	99.70	72.25	0.29
3000	99.50	72.10	0.50
2000	99.80	72.32	0.19
500	92.60	67.10	7.40
400	91.50	66.30	8.50
300	90.70	65.72	9.30
200	83.70	60.65	16.30
190	82.10	59.49	17.90
180	81.60	59.13	18.40
170	79.80	57.83	20.19
150	81.00	58.70	18.99
140	80.70	58.48	19.29
130	78.90	57.17	21.10
100	77.70	56.30	22.30

表 3.5: 次元縮小前後の判定変化（140 次元)

トピック	×→○	○→×	判定変化無し	判断付かず	総文書数
C1	0	0	100	0	7
C2	0	35.29	64.71	11.76	17
C3	9.09	0	90.9	36.36	11
C4	18.75	0	81.25	25	16
C5	0	0	100	50	8
C6	0	23.08	76.92	15.38	13
C7	0	28.57	71.43	7.14	14
C8	38.46	0	61.54	15.38	13
C9	16.67	33.33	50	8.33	12
C10	0	85.71	14.29	0	7

表 3.6: 2 グラムモデルでの各教師データトピックの総語数

トピック	C1	C2	C3	C4	C5
単語数	1029	1481	1991	2075	1672
トピック	C6	C7	C8	C9	C10
単語数	1255	710	1915	1592	1765

表 3.7: 各トピック教師データ間の共通語の出現

語数	C2	C3	C4	C5	C6	C7	C8	C9	C10	平均
C1	221	267	266	249	206	142	246	245	228	230
C2	-	352	322	320	263	186	309	321	280	294
C3	-	-	439	427	326	231	380	431	392	375
C4	-	-	-	447	354	213	407	407	381	368
C5	-	-	-	-	289	214	349	386	352	318
C6	-	-	-	-	-	178	303	301	312	274
C7	-	-	-	-	-	-	219	209	193	207
C8	-	-	-	-	-	-	-	355	348	352
C9	-	-	-	-	-	-	-	-	358	358

表 3.8: 2 グラムモデルでの正答率

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
正答数	7	3	3	9	1	15	15	8	7	1
文書数	7	19	20	15	12	15	15	15	13	7
正答率	100	15.8	15	60	8.3	100	100	53.3	53.9	14.3

表 3.9: 2 グラムモデルで C1,C6,C7 に推定された場数

推定トピック	下位 1 位	下位 2 位	下位 3 位
C1	1	92	39
C6	3	9	96
C7	111	27	0

3.5 結び

本稿では，単語分布からのトピック語モデルの検証, および次元縮小による精度変化を実験により解析し, 利用可能性を論じた．実験によって，1 グラムモデルでは単語分布によるトピック推定が平均 72% 以上の精度で可能であることを示し，さらに次元縮小率 98.59%（9923 次元から 140 次元）であっても信頼性 8 割程度の正答率を維持できることを示した．一方，2 グラムモデルでは，正答率が 50% 程度であり，トピック語モデルの検証ができなかった．この結果, 次元縮小によるトピック語モデルは 1 グラムモデルで有効であることを確認した.

表 3.10: 次元の違い

	次元数	1 度のみ出現	有効割合
2 グラム	65166	59648	8.47%
1 グラム	9923	4354	56.12%

第4章　Topic/Author 推定方式の改善

4.1　前書き

近年, インターネットが世界中に普及したことにより, 膨大な量の情報を用意に得ることができるようになった. これらの情報は一般的にテキスト形式で保持されている. これは目的に合った場所に文書として格納すべきであるが, クラスや類似性により前もって分けられていないため, 分類とクラスタリングのような機械学習手法を適用することは難しい.

実際に, “この文書は最新の提案だ”あるいは“この文書はスミス氏の手紙のようだ”ということがある. このような分類問題は情報検索 (IR) により解決する事ができる. IR では, すべての文書が単語に関するベクトルとして表わされる, ベクトル空間モデル (VSM) によって対象世界を言及する. ここで, 2 つの文書の類似性を, コサイン尺度と呼ばれる 2 つのベクトルの内積として定義する [13]. しかし, 類似性は, 共起語の発生に依存する.

しかし,IR ではテキスト形式がどう見えるか, また誰が文書を作成したかを論じるべきで, 効率に関する点だけに焦点を特化すべきでない. そのような IR は構文アプローチに基づくはずがなく, むしろ自然言語処理 (NLP) の支援が必要である.

代表例として, トピックと著者の識別がある. 著者推定問題は, 著者 (あるテキストの著者) を識別する方法を意味し, 過去一世紀以上にわたる論争の一つでもあり, テキストから何らかの特徴を捉えて著者の同定・識別を行おうとするものである. これと並んで興味あるものがトピック推定問題である. トピック (topic) とは興味ある事柄や出来事を言い, 文書を解析し何のトピックを論じるものかを推定する. これらの技術は, 文書の自動格納・自動分類や自動要約に主要な手がかりを与え, また背景や領域の推定による文脈情報を付加することで, 情報検索の効率向上に大きなヒントを与えることができる.

これまでの研究結果によると, テキストから直接有用な情報を抽出する方法では, 著者固有の性質よりもトピックとの関連性を論じるほうが分析しやすいことが知られている [16]. トピック/著者語 (T/AW) モデルとは著者推定がトピックの選定に確率分布に従うことをいう. 一方, 同一著者の下では, 各トピックは対応する語集合の多項式分布確率で表わされるとする著者語 (A/W) モデルが知られる [20]. 仮に,T/AW モデルが正しければ, 語の分布を調べることでトピック推定が可能であり, 具体的な推定手順を与える論拠となる.

しかし, テキスト形式の情報検索では高次元の処理を必要とすることから, 性能およ

び精度に差が生じる．このため精度を維持したままで効率向上を目的とした次元縮小技法がいくつか提案されている [13].

本研究では,T/AW モデルおよび AW モデルが実際に成り立つかどうかを検証する．また, 次元縮小技法を用いることによって, 効率的に処理可能であることを示す.

第 2 章はトピック・著者の語モデルおよびそれらを評価する方法を述べる．第 3 章では次元縮小, およびトピック・モデル検証にどのように技術を適用するかを述べる．第 4 章では実験によりその有用性を示し, 第 5 章では関連研究を述べ, 第 6 章で本研究の結論を述べる.

4.2 トピック, 著者, 単語のモデル化

T/AW モデルとは, 同一著者の下では, 各トピックは対応する語集合の多項式分布確率で表わされるという仮定である．つまり, 著者は, トピックに依存する多項式分布確率により語を確率的に選択する仮定されている．これはトピック, 語の分布の検証により, 文書がどのトピックを表しているかを推定することが可能であることを意味する．この状況は機械学習の分類に似ている．即ち, 訓練データを事前に準備し, どのクラスが最も適しているかを検証することに相当する．他方,AW モデルは, 著者がトピックに独立に著者自身の語集合の多項式分布確率で表されるという仮定である．つまり, 著者は多項式分布に従って語を確率的に選択する.

残念ながら, この特性を証明することはできない．一般的に,T/AW モデルは広く信じられているが,AW モデルにおいてはそうではない．本研究では, 信じられる・信じられないに関わらず, 検証を行う.

次にこれらの問題を検証する方法を述べる．根本的な問題の 1 つは語をどう扱うかである．テキスト情報は語の並びとして構成されるが, 語 (word) をどのように設定するかは自明ではない．英語では (空白などの) 特殊文字で区切られた文字列を単語と呼ぶが, 複合語 (”U.S.A.” 等のように複数の単語からなる語) や慣用句 (”get used to” 等), 連語 (”not only··· but also” 等) を考慮するかどうかは, 分析結果に大きな影響を与える．n グラム (n-gram) モデルでは, 連続する n 単語をまとめて語とみなすが, 単語の区切りを無視して数え上げるため, 多くのミスを含む可能性がある [21]．本研究では,1 グラムモデル ($n = 1$) を用いて検証を行う.

同様に考慮すべき点として内容語の問題がある．効果的に処理を行うには, 特有の意味を持つ語を選ぶ必要があろう．T/AW モデルと AW モデルでは, これらの単語の分布がテキスト文書の意味的な様相を捉えるので, 本研究では内容語だけを分析する．ここでは, 前もって不要語 (”a”, ”the” 等) を除去し, ステミング処理も行う．ここでは機能語 (”when”, ”and”等) に関しては考えない．重みには語の出現頻度 (TF) や逆出現頻度 (IDF) [1] を用いて検証を行う.

[1]IDF は n を文書数,n_i を語 w_i を含む文書数としたときの, $\frac{n}{n_i}$ の割合をいう.

与えられた 2 つの分布 (既知のトピック分布と別の 2 つの分布) をテストすることで, 実際に 2 つの分布が独立であるかどうかを調べたい. 検証する方法として本研究では 2 つの手法を用いる. 1 つ目は, カイ二乗値を用いる方法である. 2 つの分布 p,q がどのぐらい独立しているかを調べるため, 各語 w_i について, 訓練データとテストデータ 2 つの頻度分布 p_i,q_i からなる. X^2 値を以下に定義する.

$$X^2 = \sum_i \frac{(q_i - p_i)^2}{p_i} \tag{4.1}$$

2 つの分布が類似するほど, 定義より算出される X^2 値は少なくなる. テスト文書を与えられたとき, 語分布を計算し各訓練データに対する X^2 値を計算し, X^2 値が小さくなったトピックを当該のものと推定する. 2 つ目は KL ダイバージェンス $KL(p||q)$ を用いる方法である. 2 つの分布 p と q を用いることで,p から q を判別することができる. 定義より, 値が小さいときは 2 つの分布が類似していることを意味する.

$$KL(p||q) = \sum_i p_i \log \frac{p_i}{q_i} \tag{4.2}$$

本研究では推定結果の上位に正解が含まれる場合, 正しいトピックに推定ができたと考える. 実験では上位 1 位または 3 位までを扱う.

4.3 T/AW モデルと次元縮小

テキスト形式の情報検索では高次元の処理を必要とすることから, 多大な計算量を要する. これまで, 精度を維持したままで効率向上を目的とした次元縮小技法がいくつか提案されている [13].

情報検索アプローチでは, 文書 d に出現する内容語 $w_1, .., w_n$ のベクトルで表現する.

$$d = (v_1, ..., v_n)$$

ここで v_i は語 w_i に対応する数値であり一般に出現頻度であることが多い. 2 つの文書 $d1$,$d2$ が類似している場合, $d1$, $d2$ の類似度は出現数の分布を用いて定義され, 内積 $(d1, d2)$ によって与える. この方法は, モデル化が単純であり類似度も簡単に算出できることから, 広く利用されているが, 解が重み付け方法に依存し, 次元数が数万にも及ぶ高次元データをそのまま扱うと, 効率, 計算機容量の確保および即応性への対応が困難になる. 次元縮小技術を必要とするような多大な負荷 (CPU とメモリ) がかかるため, 次元縮小技法を用いることで, 高次元文書ベクトルを低次元空間に射影し, 効率よく探索範囲を絞り込むことができる.

次元縮小技法はこれまでも多数提案されてきた [13]. 主な技術として, 潜在意味索引つけ (Latent Semantic Indexing, LSI) 技法と, ランダムプロジェクション (Random Projection, RP) 技法がある. LSI は原データを用いて線形代数の理論を基にしてた特徴値を算出するため, 極めて高精度に縮小可能である. しかし, 特異値分解 (SVD) の理

論計算に多大な時間がかかり, 微小な変更でも再計算を要求することから動的な環境に利用できない.

これに対して,RP は乱数技法により次元縮小するため, 次元縮小手続きの効率が良く, 低次元空間に縮小するほど少なくて済む利点がある. さらに, 技法は文書に依存せず確率を用いるため, 動的な環境の下でテキスト文書集合が増えても再計算を要求することがない. このため, 本研究では RP を用いて次元縮小を行う. 反面, 精度が悪く適用範囲に限界がある [19].

以下では語数 d, 文書数 N とし, $X \times N$ 語・文書行列 X を $k \times N(k \ll d)$ の語・文書行列 X_{RP} に射影する. 射影を行うため,$k \times d$ の RP 行列 $R = ((r_{ij}))$ を生成する. この際, 行列 X の i 行 j 列の要素 X_{ij} は, 文書 j における語 i の頻度を意味する. 単語・文書行列 X の RP による次元縮小の計算は以下のように定義される.

$$X^{RP}_{k\times N} \quad = \quad R_{k\times d} X_{d\times N} \tag{4.3}$$

これを定義するため, 次元縮小行列 $R = ((r_{ij}))$ を, 発生確率 p に対して, 次の分布に従うように決定する [12]. $(i = 1...k, j = 1...d)$

$$r_{ij} \quad = \quad \sqrt{3} \cdot \begin{cases} +1 & (p = 1/6) \\ 0 & (p = 2/3) \\ -1 & (p = 1/6) \end{cases} \tag{4.4}$$

単語数 d, 文書数 N を k 次元に縮小する際の行列の生成に対する計算量は $O(kd)$ であり, $k \ll d$ でもあることから, 実際の処理は高速である.

$d \times N$ 行列 X から各列ベクトルを取り出し, 予め用意した訓練用データを用いて RP 行列を作成しテストデータを比較・評価する. 本研究では,RP を用いた検索では縮小率に伴う正答率の低下を評価する.

4.4 実験

この章では実験により, TA/W モデルおよび AW モデルが実際に成り立つかどうかを検証する. また,RP 技法による次元縮小の効果を調べる.

4.4.1 T/AW モデルの検証

最初の実験では T/AW モデルが成り立つかどうかを検証する. グーテンベルク・プロジェクト [18] から 3 人の著者 (Charles Dickens,George Alfred Henry and Robert Louis Stevenson,1 人当たり 10 作品) を選び, 合計 30 作品 (トピック) を選ぶ. さらに各トピックをユニットの集合に分割する. 各ユニットはそれぞれ 20 の段落から成り,1 つの文書として考える.

訓練データのユニット数が 1,2,3,4,5 のそれぞれの場合を考え, テストデータは各トピックあたり 10 ユニットを扱い実験を行う. 表 4.1 はデータを訓練する際の異語数を示している. 例えば,C.Dickens の作品 C1 の 4 ユニットでは別の 1404 語が出現する.

著者	トピック C1,..,C10
Charles Dickens	Barnaby Rudge, Bleak House, Little Dorrit, Master Humphrey's Clock, Mudfog and Other Sketches, Reprinted Pieces The Chimes, The Haunted Man and the Ghost's Bargain, The Old Curiosity Shop, The Uncommercial Traveller
George Alfred Henry	A Knight of the White Cross, Among Malay Pirates, At Agincourt, Beric the Briton, Forest and Frontiers, In Freedom's Cause, In the Reign of Terror, One of the 28th, The Bravest of the Brave, The Lion of the North
Robert Louis Stevenson	A n Inland Voyage, David Balfour (Second Part), Island Nights' Entertainments, Kidnapped, Master of Ballantrae, Memoir of Fleeming Jenkin, Merry Men, New Arabian Nights, Prince Otto (a Romance), The Black Arrow

訓練データ数	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
C.Dickens										
5	1595	2093	1565	2464	1908	2365	1411	1452	1156	2298
4	1404	1783	1243	1922	1464	1910	1194	1295	1003	1853
3	1198	1542	1076	1430	1241	1806	1028	1045	870	1332
2	1042	963	892	976	870	1099	876	768	606	1102
1	519	696	581	543	516	585	670	491	335	598
G.A.Henry										
5	1954	1506	1718	2083	2690	2216	1382	1284	1988	1554
4	1487	1060	1125	1902	2149	1709	1037	1066	1684	1008
3	1038	738	482	1515	1552	1126	817	818	1177	620
2	612	455	203	830	1070	828	521	342	639	378
1	286	269	116	353	537	436	156	154	291	218
R.L.Stevenson										
5	2799	1082	1604	1576	1775	4698	1672	1290	1452	1228
4	2263	896	1367	1419	1252	3353	1281	1081	1157	926
3	1719	784	968	1072	813	2398	960	808	993	693
2	1133	634	641	855	497	1946	455	646	771	499
1	585	196	407	462	253	774	69	279	473	292

表 4.1: 異語数

訓練データ数	Dickens	Henry	Stevenson
5	51	76	71
4	51	73	61
3	58	70	56
2	58	79	63
1	58	75	53

表 4.2: Best3 の正解率 (%)

TF,TF∗IDF を重みとして,X^2 値,KL ダイバージェンスを用いて分布を検証する. 表 4.2, 表 4.3 に上位 1 位が正解となる場合 (Best1) と上位 3 位までに正解を含む場合 (Best3) を示す.

4 ユニットの訓練データにおいて, Best3 の正解率は C.Dickens が 51%,G.A.Henry が 73%,R.L.Stevenson が 61%である. 見て分かるとおり, 訓練データが多いほど正解率は高くなっている. Best1 の場合では,5 ユニットの正解率が最も高い.

TF の代わりに TF∗IDF(+X^2 値) を扱うことで, 正解率の向上を得た. しかし,KL ダイバージェンスを用いたケースでは正解率の低下が見られた. これは, 各作品 (トピッ

訓練データ数	Dickens	Henry	Stevenson
5	35	64	48
4	39	55	43
3	44	53	44
2	40	61	47
1	28	63	35

表 4.3: Best1 の正解率 (%)

手法	Dickens	Henry	Stevenson
(Best1)			
TF*IDF + X^2	23	72	49
TF + X^2	39	55	43
TF*IDF+ KL	19	42	19
TF + KL	10	24	19
(Best3)			
TF*IDF + X^2	52	79	65
TF + X^2	51	73	61
TF*IDF + KL	33	40	39
TF + KL	32	49	30

表 4.4: 4 ユニットにおける正解率

ク) が独自の分布を持っており, KL ダイバージェンスと比べると IDF がより効果的であることが分かる.

このことより,T/AW モデルが正しく成り立っているということができる. 実際, 正解率は C.Dickens が 58%,G.A.Henry が 79%,R.L.Stevenson が 71%(Best3+ X^2) となった. TF∗IDF を用いることにより, 正解率を改善し, より多くのデータにおいて精度の向上が見られた.

4.4.2 AW モデルの検証

第 2 の実験では AW モデルが実際に成り立つかどうかを検証する. 実験データは実験 1 と同じコーパスを用いる. テストのため, 各トピックを再びユニットに分割する. 各ユニットは 20 の段落から成る. 今回の実験では訓練データは存在しない. コーパスを通して得ることができる各著者の作品に表れた語分布を用いて実験を行う. ここでは, 著者推定と著者の語分布の 2 つの特性を調べる.

表 4.5 にコーパスの語分布を示す. 明らかに, 分布の中では多くの単語が共通して出現しており, 著者の間で明確に区別する事ができない (50%〜70%).

初めに著者間の分布がどれぐらい類似しているかを検証する. 前述のとおり, 著者のすべての作品の単語頻度を数えることによって (著者ごとに) すべての語分布を取る. さらにすべての著者に対して, 著者のすべての作品からそれぞれ 5 つのテストユニットの分布, つまり 150 の (=3 × 10 × 6) の分布を取る. その後, 各著者の分布と比較して X^2 値を計算し,χ^2 検定を適用する. 5 つのテスト・ユニットに関しては著者を識別するために平均した X^2 値を用いる. この結果を表 4.6 に示す.

	C.Dickens	G.A.Henry	R.L.Stevenson
C.Dickens	16895	8699	10014
G.A.Henry	-	12548	8169
R.L.Stevenson	-	-	15764

表 4.5: 著者間の共通語数

	C.Dickens	G.A.Henry	R.L.Stevenson
総単語数	661207	407558	316443
異語数	16895	12548	15764
χ^2 (99.5%)	13758.0	10215.2	12836.3
χ^2 (0.5%)	20378.2	15140.0	19015.3
(著者の分布との比較)			
By C.Dickens	548728	770757	590290
By G.A.Henry	583830	735636	576884
By R.L.Stevenson	609703	703688	559996
(5 ユニットの平均)			
By C.Dickens	647934	398542	305655
By G.A.Henry	648115	397556	305921
By R.L.Stevenson	6648002	398592	304928

表 4.6: χ^2 検定の X^2 値

結果より, 語分布が著者自体に依存する訳ではないことが分かる. X^2 値はすべて χ^2 値を越えており, 分布が著者を推定するとは断定できないということができる. 信頼性 0.5%の場合でさえも充足する X^2 値は存在しない. 最小となる X^2 値は R.L.Stevenson 以外では間違った著者に現れる. さらに悪いことには, すべての場合において 5 つのユニットを平均したものが非常に悪くなってしまう.

第 2 の問題はテストデータの著者をどのぐらい識別する事ができるのかという問題である. 実験の目的は異なるが, 表 6 を用いて X^2 値を計算する. また著者を識別する際には Best1 評価方法を適用する. 結果を表 4.7 に示す. 作品全体および 5 ユニットの平均における正解率は 63.3%および 33.3%である. これを見ると結果が良いようにも見えるが, ほとんどが R.L.Stevenson に推定されている. 実際 5 ユニットの場合はすべて R.L.Stevenson に推定されている.

	C.Dickens	G.A.Henry	R.L.Stevenson	Total
(著者の分布との比較)				
C.Dickens	4	0	6	40%
G.A.Henry	0	5	5	50%
R.L.Stevenson	0	0	10	100%
(Average)				63.3%
(5 ユニット)				
C.Dickens	0	0	10	0%
G.A.Henry	0	0	10	0%
R.L.Stevenson	0	0	10	100%
(Average)				33.3%

表 4.7: 著者推定の正解率

表 4.8 からは, 作品全体の場合に関して異語が出現していることが分かる. 同時に

	異語数	総単語数	推定先の著者
(C.Dickens)			
C1	7765	112758	C.Dickens*
C2	9059	141905	C.Dickens*
C3	8997	139133	C.Dickens*
C4	4073	22061	R.L.Stevenson
C5	3588	12905	R.L.Stevenson
C6	6703	42035	R.L.Stevenson
C7	2910	13025	R.L.Stevenson
C8	3009	13096	R.L.Stevenson
C9	7386	97455	C.Dickens*
C10	8555	67232	R.L.Stevenson
(G.A.Henry)			
C1	4786	53533	G.A.Henry*
C2	3589	25899	R.L.Stevenson
C3	4389	45303	G.A.Henry*
C4	4728	55643	G.A.Henry*
C5	3586	14904	R.L.Stevenson
C6	4795	48147	R.L.Stevenson
C7	3891	32714	R.L.Stevenson
C8	4388	47130	G.A.Henry*
C9	4282	37170	R.L.Stevenson
C10	4736	47445	G.A.Henry*
(R.L.Stevenson)			
C1	4063	16170	R.L.Stevenson*
C2	5298	36317	R.L.Stevenson*
C3	3025	20663	R.L.Stevenson*
C4	4506	30502	R.L.Stevenson*
C5	4970	40326	R.L.Stevenson*
C6	5509	25577	R.L.Stevenson*
C7	6000	40295	R.L.Stevenson*
C8	5903	42263	R.L.Stevenson*
C9	4833	30187	R.L.Stevenson*
C10	4822	34831	R.L.Stevenson*

表 4.8: 著者推定の際の異語数と総単語数

C.Dickens と G.A.Henry において, 4 つあるいは 5 つが正確に推定されていることもわかる. これは異語数が推定に影響したことを意味する. 比較的多い異語が R.L.Stevenson の分布との違いを表し, X^2 値を小さくしたと考えられる. これは特定の著者に対して偏った分布を持つことも意味している.

著者推定に関しては正解率 60%を超えたが, これは全体の単語量が推定に影響をしたことによるもので, 必ずしも分布の違いから得られた結果ではない. 従って AW モデルに関して, それを信頼するだけの理由と考えることはできない.

4.4.3 次元縮小

最後の実験は RP の適用精度に関するものである. 実験データとして,W.Shakespeare の作品より 10 作品を選ぶ [15]. 各作品をトピックとし, 各トピックでの全ての第 1 章を訓練データとして使用する. データはすべてステミング処理と不要語除去した後,10 トピックの第 1 章の語分布を調べる. その後, 他の章の全 139 場面に関してもステミング処理と不要語除去を行い, 語分布を抽出する. この抽出した分布を訓練データから得ら

れた分布と比較することで, 場面がどのトピックに属するのかを識別する. 表 4.9 に前処理を行った後の訓練データの単語数を示す.

トピック	作品名	章/場面
C1	A Midsummer Night's Dream	5/9
C2	As You Like It	5/22
C3	Cymbeline	5/26
C4	Hamlet	5/20
C5	Othello	5/15
C6	Julius Caesar	5 /18
C7	King John	5/16
C8	Richard II	5/19
C9	Henry VIII	5/17
C10	The Tempest	5/9

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
654	783	1171	1327	1132	849	503	1120	1138	1067

表 4.9: 訓練データの異語数

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
100.0	100.0	40.0	33.3	75.0	86.7	100.0	46.7	76.9	100.0

表 4.10: 正解率

最初に T/W モデルが成り立つかどうかを再度検証する. 表 4.10 に実験結果を示す. 実験では 72.46%の正解率を得た (Best3) ため T/W モデルが成立していると言ってよい. RP 手法による結果は任意に生成された RP 行列に依存する. 本稿においては 10 回の実験の平均値を評価する. 表 4.11 に結果を示す. 表は縮小された次元 (次元) と正解率 (正確さ) の関係を示す. "精度低下"は次元を縮小したことによる正解率の低下割合を意味する.

表より,C3 と C4 で低い正解率となっていることが分かり,C3 と C4 には多くの単語が出現している. また表 4.12 は 2 つのトピックのデータ間の共通単語を示している. 表 4.12 より,C3 も C4 も他と比べて明確な違いは存在しないことが分かる. しかし, 僅かではあるが他と比べるとこの 2 つのトピック間においては共通の単語が多い. これは C8 にも共通する.

表 4.11 より,500 次元で 67%の正解率,140 次元 (98.59%の縮小率) では正解率の低下が 19.30%となっていることが分かる. これらは RP 手法を用いた次元縮小が非常によく働いていることを意味する. 表 4.13 は 140 次元に次元を縮小した場合にトピックの識別判断がどう変化したかを示している. 表の「NO/YES」は不正解だったものが 140 次元に縮小する事で正解と判断されたことを意味し, 「Yes/NO」は正解だったものが不正解と判断されたことを意味する. 「NoChange」は識別されたトピックの変更はなかったことを意味し, 「UnKnown」は判断できない (Yes と No が 50%ずつ) ことを意味する. 識別が変更されたトピックはほとんどないことから, 次元を縮小することで T/W モデルの充足性は変わらない.

次元数	正解率	精度低下
9923	72.46	0.0 %
9000	71.09	1.89
5000	72.25	0.29
3000	72.10	0.50
2000	72.32	0.19
500	67.10	7.40
400	66.30	8.50
300	65.72	9.30
200	60.65	16.30
190	59.49	17.90
180	59.13	18.40
170	57.83	20.19
150	58.70	18.99
140	58.48	19.29
130	57.17	21.10
100	56.30	22.30

表 4.11: 次元縮小と正解率

	C2	C3	C4	C5	C6	C7	C8	C9	C10
C1	221	267	266	249	206	142	246	245	228
C2	-	352	322	320	263	186	309	321	280
C3	-	-	439	427	326	231	380	431	392
C4	-	-	-	447	354	213	407	407	381
C5	-	-	-	-	289	214	349	386	352
C6	-	-	-	-	-	178	303	301	312
C7	-	-	-	-	-	-	219	209	193
C8	-	-	-	-	-	-	-	355	348
C9	-	-	-	-	-	-	-	-	358

表 4.12: 各トピック訓練データ間の共通語数

トピック	NO/YES	YES/NO	NoChange	Unknown
C1	0	0	100	0
C2	0	35.29	64.71	11.76
C3	9.09	0	90.9	36.36
C4	18.75	0	81.25	25
C5	0	0	100	50
C6	0	23.08	76.92	15.38
C7	0	28.57	71.43	7.14
C8	38.46	0	61.54	15.38
C9	16.67	33.33	50	8.33
C10	0	85.71	14.29	0

表 4.13: 次元縮小前後の判定変化 (140 次元)

4.5 関連研究

原作者の問題とは, テキストや他の特徴をを調べることによってどのように著者を識別するかを意味する. 応用例として, シェークスピアが実際に生きていたかどうか, 日本のグリコ森永事件における脅迫状の分析が代表的である. グリコ・森永事件とは, 日本の産業製菓業江崎グリコおよび森永に主として向けられた, 恐喝事件で, 現在未解決のままとなっている [14]. この事件は容疑者,「怪人 21 面相」として知られている個人またはグループとのやり取りが, グリコの社長が誘拐されてから最後に接触するまで, 全体で 17ヶ月かかった事件である. 詳しくは http://ja.wikipedia.org/wiki/グリコ・森永事件を参照.

著者推定・分析を行うためには, 文体の計量的特長 (stylometry), 例えば語長・文長・語数や機能語 (while, on などの不要語記号) などを調べる方法があるが, 同一筆者でも差が大きく特徴が有効とはいいがたい [16].

トピック推定は, 効率よく検索するための文脈に依存した情報の評価や要約, トピックへの文書の自動分類の仕方とも関係がある.

同一著者の下では, 各トピックは対応する語集合の多項式分布確率で表わされるとする T/W モデルが議論されることが多い [20]. これが正しければ, トピック上の語分布を検討および確率分布の識別をすることで, トピック推定が可能になる. 一般に, 文書は複数トピックを含むが, 本研究では文書とトピックを同一視し, トピック推定を効率よく実現する手法を考える. 代表例はニュース記事, つまりニュース放送の翻訳である.

4.6 結び

本研究では,T/AW モデルが経験的に適用できることを示した. また,TF∗IDF は X^2 分布分析と併用する事で上手く働くことを述べた. さらに,AW モデルを実際に適用することはできず, 著者推定ではなくトピック推定を考えることが得策であることを示した. 次にトピック推定に適した次元縮小を提案しランダム・プロジェクションが実際に有用であることを確認した. 実験結果より 1 グラムモデルにおいてトピック語モデルを仮定することができ, RP 技術が次元数の 98.59%を縮小しても良い有効性を維持することができることを示した.

第5章　結論

本研究では, テキスト形式で保持されてデータの管理・検索をより高度に行うための支援の一つとして, テキストデータのトピックを推定する手法とその改善について論じた.

まず, 共起語を考慮に入れた EM アルゴリズムによるテキスト分類の新たな手法を提案した. 実験の結果から, 共起語の概念を取り入れることで分類精度の向上を図ることに成功した. またこの際, 共起語の追加に高いしきい値を設けることで, しきい値を低く設定した場合よりも, より相関的な共起語を考慮に入れることで, テキスト分類の精度を向上させることができる.

次に,T/AW モデルの検証および推定方法の改善を提案した. 検証の結果,T/AW モデルを用い, 著者の作品の語分布を調べることで, トピックの推定を行うことができた. また推定方法の改善として, いくつかの実験から, 単語の重みとして従来の頻度ではなく,TF*IDF 値を用いることで精度の向上を図ることができた.

また,T/AW モデルにランダムプロジェクションを適用し, モデルに対する効率的で有効な処理を行った. ランダムプロジェクションを適用する事で, 次元縮小率 98.59%であっても信頼性 8 割の正答率を維持することが可能であり, 効率, 計算機容量の確保及び即応性への対応が可能となった.

これにより, トピックを推定する事で, テキストデータの自動分類が可能となった. またこれは, 作品には固有の特徴, 語の分布が存在していると言うことができる. テキストデータの特徴を上手く捉えることが, トピックの推定に必要である. この際, モデルに対してランダムプロジェクション技法を適用することで, モデルを崩さずに次元を縮小することができ, 次元数が数万におよぶ高次元データに対しても, 効率の良さを確保しつつ, 同様の手法で情報へのアクセスを支援することができるようになると言える.

謝辞

本研究を遂行するにあたり，日頃より数々のご指導をいただいた，法政大学工学部情報電気電子工学科 三浦孝夫教授に深く御礼申し上げます.

また，産能大学経営情報学科 塩谷勇教授にも多くのご指導をいただきました．深く感謝いたします.

データ工学研究室の先輩方，同輩，後輩たちにも，本研究の遂行にあたって数多くの助言と快適な研究環境の整備をして頂きました．御礼申し上げます.

修士論文として私の研究をまとめることができたのも，多くの皆様方の御支援，御協力の賜物であります．この場をお借りしまして，厚く御礼申し上げます.

最後に，今までの学生生活を支えてくださった私の両親に感謝したいと思います.

参考文献

[1] J.Han, et.al : Data Mining: Concepts and Techniques, Morgan Kaufmann Pub., 2000

[2] 岩崎 学: 不完全データの統計解析, エコノミクス社,2002

[3] 松尾, 石塚: 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム, 人工知能学会論文誌 17-3-D, pp.217-223, 2002

[4] T.Mitchell: Machine Learning, McGraw-Hill Education, 1997

[5] Nigam, K., McCallum, A.K., Thrun, S. and Mitchell, T.M.: Text Classification from Labeled and Unlabeled Documents using EM , Machine Learning Vol .39, No.2, pp. 103–134, 2000

[6] 大澤 幸夫, ネルス E. ベンソン, 谷内田雅彦,: KeyGraph : 語の共起グラフの分割・統合によるキーワード抽出, 電子情報通信学会論文誌　D-I Vol.J82-D-I No.2 pp.391-400, 1999

[7] Poter, M.F.:An algorithm for shuffix stripping , Program, Vol. 14, No. 3, pp.130-137,1980

[8] 上嶋 宏, 三浦 孝夫, 塩谷 勇,: Estimating Timestamp From Incomplete News Corpus, Journal of Communications in Information and Systems : Special Issue on Computational Informatics in Data Mining and Information Retrieval, Vo.4, No.4, International Press, pp.273-288 , 2005

[9] 上嶋 宏, 三浦 孝夫, 塩谷 勇,: Improving Text Categorization by Synonym and Polysemy, SYSTEMS AND COMPUTERS IN JAPAN, Vol. 36-4, pp.1-8, 2005 April

[10] 吉原 幸輝, 三浦 孝夫, 塩谷 勇: Classifying Melodies by Using EM Algorithm, IEEE Computer Software and Application Conference (COMPSAC), pp.204-210, 2005

[11] 新納 浩幸, 佐々木 捻.: EM アルゴリズムの最適ループ回数の予測を用いた語義判別規則の教師なし学習，情報処理学会論文誌 Vol.44,No.12,2003

[12] Achloiptas, D.: Database-friendly random projections, ACM-PODS 2001, pp.274-281

[13] 北研二，他: 情報検索アルゴリズム，共立出版， 2002

[14] 村上征勝：シェークスピアは誰ですか?–計量文献学の世界，文藝春秋社， 2004

[15] The Complete Works of William Shakespeare, http://shakespeare.mit.edu/works.html

[16] E.Stamatos, N.Fakotakis, G.Kokkinakis: Automatic Authorship Attribution, EACL, 1999

[17] M.Steyvers, P.Smyth, T.Griffiths : Probabilistic Author-Topic Models for Information Discovery，KDD， 2004

[18] Gutenberg Project, http://www.gutenberg.org

[19] Oh'uchi, H., Miura, T. and Shioya, I.: Document Retrieval using Projection by Frequency Distribution, Intn'l J. on Artificial Intelligence Tools (IJAITS), Special Issue, Vol.16, 2007

[20] M.Steyvers, P.Smyth, T.Griffiths : Probabilistic Author-Topic Models for Information Discovery, KDD, 2004

[21] Nakayama,M., Miura, T.: Identifying Topics by using Word Distribution, IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), 2007,pp.245-248