

Topic/Author 推定方式の改善

中山 基[†] 三浦 孝夫[†] 塩谷 勇^{††}

[†] 法政大学 工学研究科 電気工学専攻 〒184-8584 東京都小金井市梶野町 3-7-2

^{††} 産能大学 経営情報学部 〒259-1197 神奈川県伊勢原市上粕屋 1573

E-mail: †{c02d3076,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

あらまし 本稿では、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。また、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。次に、次元縮小技術の1つであるランダム・プロジェクトを用いて、著者、トピックを推定するための効率的な処理方法を示す。最後に実験結果を示し有効性を示す。

キーワード データマイニング, 知識発見, テキストマイニング

Topic/Author Improvement of the Identification technique

Motoi NAKAYAMA[†], Takao MIURA[†], and Isamu SHIOYA^{††}

[†] Dept.of Elect.& Elect. Engr., HOSEI University 3-7-2, KajinoCho, Koganei, Tokyo, 184-8584 Japan

^{††} Department of Management and Information Science, SANNO University 1573, Kamikasuya, Isehara city, Kanagawa 259 1197 Japan

E-mail: †{c02d3076,miurat}@k.hosei.ac.jp, ††shioya@mi.sanno.ac.jp

Abstract In this work, we examine and verify Topic Author Word model which says each topic can be identified by means of word distribution under same author, and also we examine Author Word model to see it is unsuitable for the identification. Then, by using Random Projection which is one of the dimension reduction techniques, we show we can obtain efficient and effective processing to identify topic/author. We show some experimental results to see the effectiveness.

Key words data mining, knowledge discovery, text mining

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、膨大な量の情報を容易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、全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 [2]。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 [7]。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 (AW) モデルが知られる [8]。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 [2]。

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进行评估する方法を述べる。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。

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と仮定されている。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 [9]。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る。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)

(注1) を用いて検証を行う。

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。

$$X^2 = \sum_i \frac{(q_i - p_i)^2}{p_i} \quad (1)$$

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して推定する。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。

$$KL(p||q) = \sum_i p_i \log \frac{p_i}{q_i} \quad (2)$$

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。

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 [2]。

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

$$d = (v_1, \dots, v_n)$$

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 d_1, d_2 が類似している場合、 d_1, d_2 の類似度は出現数の分布を用いて定義され、内積 (d_1, d_2) 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 [2]。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)

(注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。

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とも済む利点がある。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 [5]。

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定義される。

$$X_{k \times N}^{RP} = R_{k \times d} X_{d \times N} \quad (3)$$

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 [1]。 ($i = 1 \dots k, j = 1 \dots d$)

$$r_{ij} = \sqrt{3} \cdot \begin{cases} +1 & (p = 1/6) \\ 0 & (p = 2/3) \\ -1 & (p = 1/6) \end{cases} \quad (4)$$

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. 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。

4.1 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。ゲーテンベルク・プロジェクト [3] から 3 人の著者 (Charles Dickens, George Alfred Henry and Robert Louis Stevenson, 1 人あたり 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 ユニットの扱い実験を行う。表 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 ユニットの別々の 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 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 2 Best3 の正解率 (%)

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 3 Best1 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 2, 表 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。

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 ユニットにおける正解率

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 TF*IDF (+ X^2 値) 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出来る。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 と X^2 値を組み合わせる用いることにより,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。

4.2 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 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,著者間で明確に区別する事ができない (50%~70%)。

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

表 5 著者間の共通語数

初めに著者間の分布がどれくらい類似しているかを検証する。前述のとおり,著者の全ての作品の単語頻度を数えることによって(著者ごとに)全ての語分布を取る。さらに全ての著者に対して,著者の全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 ($=3 \times 10 \times 6$) 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 6 に示す。

	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 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 値は全て 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出来る。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,全ての場合において 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ふ。

第 2 の問題はテストデータの著者をどのく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 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 ユニットの場合は全て 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 ユニットの)				
C.Dickens	0	0	10	0%
G.A.Henry	0	0	10	0%
R.L.Stevenson	0	0	10	100%
(Average)				33.3%

表 7 著者推定の正解率

	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 8 著者推定の際の異語数と総単語数

表 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 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。

4.3 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 [6]。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 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 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 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 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 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 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 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 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 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 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 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 13 次元縮小前後の判定変化 (140 次元)

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 [4]。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えない [7]。

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 [8]。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。

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。

文 献

- [1] Achloiptas, D.: Database-friendly random projections, ACM-PODS 2001, pp.274-281
- [2] 北研二, 他: 情報検索アルゴリズム, 共立出版, 2002
- [3] Gutenberg Project, <http://www.gutenberg.org>
- [4] 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
- [5] 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
- [6] The Complete Works of William Shakespeare, <http://shakespeare.mit.edu/works.html>
- [7] E.Stamatos, N.Fakotakis, G.Kokkinakis: Automatic Authorship Attribution, EACL, 1999
- [8] M.Steyvers, P.Smyth, T.Griffiths : Probabilistic Author-Topic Models for Information Discovery, KDD, 2004
- [9] 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