社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－DBS－142 (1)
2007－ＦＩ－ 87 (1)
2007／5／31

単語分布からのトピック推定

中山 基　　三浦 孝夫

法政大学 工学研究科 電気工学専攻
東京都小金井市梶野町 3-7-2

概要

この論文では，トピック語モデルを検証する．トピック語モデルとは，同一著者の下での単語分布により，各トピックを識別することができるモデルである．また，次元縮小手法の 1 つであるランダムプロジェクション手法を用いることで，モデルへの効率的で有効な処理の方法を示す．本稿では，シェークスピア作品を検査し，それらの戯曲の各場面を正確に識別することができることを示す．

Identifying Topics by using Word Distribution

Motoi NAKAYAMA　　Takao MIURA

ept.of Elect. and Elect. Engineering, Hosei University
Kajinocho 3-7-2, Koganei, Tokyo, Japan

Abstract

In this work, we examine and verify a *topic word model* which says each topic can be identified by means of word distribution under same author, and by using Random Projection, one of the dimension reduction techniques, we show we can obtain efficient and effective processing to the model. We examine Shakespeare works and show we can identify scenes correctly to their dramas.

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 [3]．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 [5]．

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 [5]．著者トピックモデル (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 [6]．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 (IR) 分野の研究活動では、全ての文書が言葉に関するベクトルとして表わされる場合、ベクトル空間モデル (VSM) による利益の世界について記述することは一般的で、2 つの文書の類似性を、2 つのベクトルの内積として容易に計算することができる．[2] しかし, 類似性は単語頻度 (TF) か, 逆の文書頻度 (IDF) のような単語にどのような重み付け方法を用いるかに依存し, 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 [2].

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．

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，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．この考え方は, 機械有り学習と似ており, 事前に訓練データから特徴を抽出して, この結果に最も似たものを選べばよい．

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 等のような語) や共起性の強い語 (”get used to”のように一緒に使われることで違う意味となる語) 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．

分布の独立性を調べるため, 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検定を行う．しかし, 語の出現が疎であるため単純な適用が難しいこともある．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．

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

上式では，X^2 値が小さいほど分布が類似することを表している．また正答率は，検定文書中の正しく推定された文書数を p，文書総数 (総場面数) を q とするとき次式で定義される．

$$\frac{p}{q} \quad (2)$$

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 [2]．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 が類似すれば出現数の分布も類似する．このため, 類似度を内積 (d_1, d_2) によって表せばよい．

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 [2]. 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.

次元縮小技法には 2 つの手法が代表的である. 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が必要とな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，テキスト文書集合に独立に縮小できるため, 変更が生じ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，テスト (未知) 文書に対する推定を行うため 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 の頻度である. 語・文書行列 X の RP による次元縮小の計算とは以下のように定義される[1]:

$$X^{RP}_{k\times N} = R_{k\times d} X_{d\times N} \tag{3}$$

これを定理するため, 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]. $(i = 1...k, j = 1...d)$

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

こ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.

訓練文書の $d \times N$ 行列 X から各列ベクトルを取り出し，テスト文書ベクトルデータを用いて要素値を求めて比較・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.

[1]データの初期生成に対する計算量は $O(dkN)$ となる.

4 実験

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 [4]. 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.

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

ここで, トピックとしの各作品第 1 章の全ての場から単語を抽出し，各トピックにおける教師データとする. これらのデータ全てに，ステミング, 不要語処理を行った後, 10 の作品について単語分布を検証する. 表 1 に, 各トピック教師データの単語数を示す.

また,2 章以降の全 138 場に対してもステミング, 不要語処理を行い, 単語を抽出して各場を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，教師データの単語分布と比較することでトピックを推定する.

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.

なお, 本稿では,X^2 値の観点から，ランキング形式でベスト 3 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.

4.2 実験結果

実験結果を示す.

表 1 に実験 1 の正答率を示す．表 2 に各文書における本来の正答トピックと, 推定結果が正答か不正答かを示す．("○"は正答を意味し，"×"は不正答を示す．)

実験 2 では次元縮小による精度低下を評価する．表 3 に縮小した次元とそのときの正答率を示す．"精度低下"は次元縮小をすることでどの程度精度が落ちたかの割合を示す.

最後に，表 4 で 2 グラムモデル分布を調べるため, 教師データの語総数と推定の正答率を示す.

表 1: 1 グラムモデルでの正答率と各単語数

トピック	単語数	正答率 %
C1	654	100.0
C2	783	100.0
C3	1171	40.0
C4	1327	33.3
C5	1132	75.0
C6	849	86.7
C7	503	100.0
C8	1120	46.7
C9	1138	76.9
C10	1067	100.0
(average)		72.46

4.3 考察（実験 1）

表 1 より, 正答率が 72.46%を得たことから, 本実験で用いるデータがトピック語モデルに従うことがわかる．表 1 を詳細に見ると，C3，C4 の正答率が低いことがわかる．C3，C4 に関しては多くの単語を含んでいる．表 6 は各トピック教師データ間において，共通に出現する単語総数を示している．これより共通に出現する語の分布は偏っておらず，C3，C4 が特殊な分布でないことが確認できる．これは C8 においても同様に考えられる.

4.4 考察（実験 2）

表 3 からわかるように，500 次元への縮小で 67%の正答率，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．ここから RP が次元縮小において適切に働いていることがわかる．表 5 では, 縮小前後における推定先の変化の割合を示している．表 5 より，変化なしの割合は

表 2: 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: 次元縮小における精度の変化

次元数	正答率	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: 2 グラムモデルでの正答率と各単語数

トピック	単語数	正答率 %
C1	1029	100.0
C2	1481	15.8
C3	1991	15.0
C4	2075	60.0
C5	1672	8.3
C6	1255	100.0
C7	710	100.0
C8	1915	53.3
C9	1592	53.9
C10	1765	14.3

表 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
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 6: 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
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	3528
C6	-	-	-	-	-	178	303	301	312
C7	-	-	-	-	-	-	219	209	193
C8	-	-	-	-	-	-	-	355	348
C9	-	-	-	-	-	-	-	-	358

高く，次元の縮小によるトピック語モデルへの影響がほとんどないことを示している．

4.5 考察（実験 3）

表 4 からわかるように，トピック C1，C6，C7 を除いて表 1 と比べて正答率が低い．しかし，表 7 からわかるように，特にトピック C1，C6，C7 には数多くの場が推定されている．また，7 からわかるようにトピック C1，C6，C7 の総単語数は他に比べて少ない．表 8 は，1 グラムモデルと 2 グラムモデルの次元と，1 度だけ出現する単語の割合を示す．1 グラムモデルでは各トピックの有効な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．X^2 値は出現頻度によって定まるため，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 グラムモデルではトピック語モデルが検証できた，とはいい難い．

表 7: 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

表 8: 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%

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いことを示した．

参考文献

[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] 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

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

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

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