DEIM2021 A24-3

固有名詞とPCAを用いたギブスサンプリングによる代表文書の選定

佐藤　謙仕† 　三浦　孝夫†

† 法政大学 理工学部創生科学科　〒 184–8584 東京都小金井市梶野町 3-7-2
E-mail: †kenji.sato.9a@stu.hosei.ac.jp, ††miurat@hosei.ac.jp

あらまし　情報検索の分野においてベクトル空間モデルは多岐に渡って利用されている．その中で，重要度として単語の出現頻度や TF*IDF を使うことが主流である．本研究では，トピック文書の抽出方法として，固有名詞の出現頻度で表した文書行列を主成分分析し，主成分得点を重要度として扱ったギブスサンプリング (GS) を提案する．主成分分析による出現単語の相関性を用いた主成分を扱うことにより，単語の潜在的な意味を扱える．また，ギブスサンプリングによるサンプル列の生成により文書行列の主成分の分布を近似することができる．

キーワード　自然言語処理，情報検索，主成分分析，　ギブスサンプリング，マルコフ遷移

1. 前　書　き

新聞やニュースには，スポーツや経済といった大枠のカテゴリがあり，それぞれのカテゴリに対して時期特有のトピックがある．例えば，2016 年の 8 月のスポーツ記事のトピックはリオ五輪である．

本研究では，そうしたカテゴリのトピックとなる代表文書を選定する方法として，固有名詞と主成分分析を用いたギブスサンプリング法を提案する．カテゴリの特徴を表した文書を代表文書とする．

情報検索の分野では用途によって名詞や形容詞，出現頻度や TF*IDF など様々な単語や重要度が扱われる．本稿ではまず，第 2 章でトピックの抽出に固有名詞と主成分分析を用いた重要度を扱うことを提案する．次に 3 章でギブスサンプリングについて述べ，4 章でギブスサンプリングを用いた代表文書の選定法を提案，5 章で実験結果を示し，6 章で結論を述べる．
ギブスサンプリングでは分布を近似するので，本稿では分布を文書の特徴として扱う．

2. 固有名詞と主成分分析を用いた重要度

トピックの抽出において，どのような単語を扱うかというのは考慮すべき問題である．そこで，今回は固有名詞を扱うことを提案する．固有名詞はトピックや文書の特徴を表す単語である．例えば，トランプ，ホワイトハウスという固有名詞が出現する記事であれば米政権に関する記事だと推測できる．このように，固有名詞を扱うことによって一般名詞や形容詞に比べて直接的にトピックを推定できる．

しかし，固有名詞を扱ううえで低頻度語が多いという問題点がある．低頻度語を扱うと文書同士の類似性を図ることが困難である．そのため，潜在的には同じ意味の文書でも全く異なる文書であると判別してしまう．

この問題を解決するために，主成分分析を用いた重要度を提案する．主成分分析は，多次元データを大きな情報の損失をせずに，低次元の指標で表す手法である．相関行列を用いた主成分分析では，相関のある多数の説明変数から相関のない少数の合成変数 (主成分) を生成する．

固有名詞の出現頻度で表した文書行列を主成分分析することにより，説明変数である固有名詞から，固有名詞の出現文書の相関性に基づいた主成分 (以下固有名詞成分) を生成し，各成分の重みを主成分得点で表すことができる．これは，出現文書が類似した単語は潜在的に類似した意味の単語であるという概念に基づいている．このことから，固有名詞成分は固有名詞の出現文書の相関性から生成されるので，潜在的な意味を表した成分であると解釈できる．これにより，単語の潜在的な意味合いを失わずに低頻度の単語を扱うことができる．

3. ギブスサンプリング

ここでは，ギブスサンプリングについての基本的な概念とアルゴリズムについて述べる．

ギブスサンプリングはマルコフ連鎖モンテカルロ (MCMC) 法の一つで，多数の確率変数の同時分布からサンプリングする手法である．ギブスサンプリングが有効なのは，サンプリングしたい確率変数が多数存在する場合，すなわち 2 次元以上の確率変数ベクトルの場合である．同時分布からのサンプリングは困難だが，それらの間の条件付き分布からのサンプリングが容易である場合に効果的である．

以下図 1 にアルゴリズムを示す．$X_{\setminus l}$ は X から x_l を取り除いたもので，m は X の確率変数の総数を表す．

l は $1, ..., m$ と変えながら周辺確率 $p(x_i|X_{\setminus l}^{(i)})$ からサンプリングを行い，$\tilde{x_l}$ を得る．それと $X_{\setminus l}^{(i)}$ を繋げたものを $X^{(t+1)}$ とする．このように，様々な周辺確率からサンプリングすることによって，同時分布を近似するサンプル列を生成することができる．
また，ギブスサンプリングにはマルコフ性があり，X は $X^{(t-1)}$ の状態のみによって決まる．遠い過去の状態から影響を受けないわけではないが，必ず直近の状態を通してのみ伝わってくる．

4. GS を用いた代表文書の選定

本稿では，3 章で述べたギブスサンプリングを用いた代表文書の選定法を提案する．代表文書とはカテゴリの特徴を表した

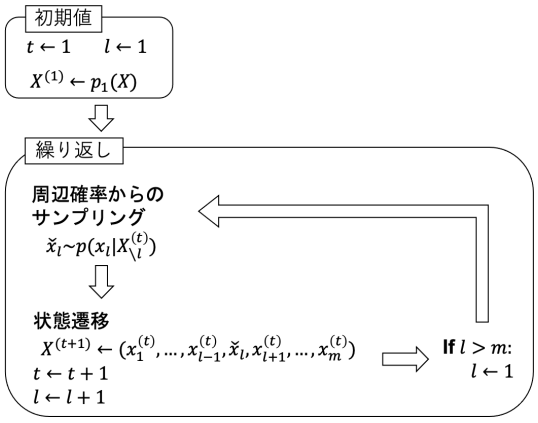

図 1　GS アルゴリズム

トピック文書であり，本研究では固有名詞成分の分布を文書の特徴として扱う．したがって，代表文書を以下の 2 つの条件で定義する．1 つ目は，カテゴリ内の代表的な固有名詞成分が文書の特徴となる固有名詞成分であること，2 つ目は，文書の内容が時期特有の旬の話題であることとする．以下の手順をカテゴリ毎に行う．

まず，ギブスサンプリングを用いた代表文書の固有名詞成分 (pc) ベクトルの生成法について説明する．

1. 訓練ベクトルの初期値 $(pc_1^{(0)}, pc_2^{(0)}, ..., pc_m^{(0)})$ を適当な分布から発生させる．

2. $pc_1^{(1)}$ を条件付き確率分布 $\pi(pc_1 | pc_2^{(0)}, ..., pc_m^{(0)})$ から発生させる．

3. $pc_2^{(1)}$ を $\pi(pc_2 | pc_1^{(1)}, pc_3^{(0)}, ..., pc_m^{(0)})$ から発生させる．

4. $pc_3^{(1)}$ を $\pi(pc_3 | pc_1^{(1)}, pc_2^{(1)}, pc_4^{(0)}, ..., pc_m^{(0)})$ から発生させる．

5. 一般に $pc^{(i)} = (pc_1^{(i)}, pc_2^{(i)}, ..., pc_m^{(i)})$ が得られたら，

(1) $pc_1^{(i+1)}$ を $\pi(pc_1 | pc_2^{(i)}, ..., pc_m^{(i)})$ から発生させる．

(2) $pc_2^{(i+1)}$ を $\pi(pc_2 | pc_1^{(i+1)}, pc_3^{(i)}, ..., pc_m^{(i)})$ から発生させる．

(3) $pc_3^{(i+1)}$ を $\pi(pc_3 | pc_1^{(i+1)}, pc_2^{(i+1)}, pc_4^{(i)}, ..., pc_m^{(i)})$ から発生させる．

というように $pc^{(i+1)} = (pc_1^{(i+1)}, pc_2^{(i+1)}, ..., pc_m^{(i+1)})$ を得る作業を $(i = 1, 2, 3, ...,)$ と繰り返す．

訓練ベクトルの固有名詞成分の分布が安定 (収束) するまで繰り返す．

次に，条件付き確率分布の選定について説明する．

$pc^{(i)} = (pc_1^{(i)}, pc_2^{(i)}, ..., pc_m^{(i)})$ において，$pc_1^{(i+1)}$ を条件付き確率分布 $\pi(pc_1 | pc_2^{(i)}, ..., pc_m^{(i)})$ から発生させる場合，条件付き確率分布 $\pi(pc_1 | pc_2^{(i)}, ..., pc_m^{(i)})$ を文書集合内で訓練ベクトル $(pc_2^{(i)}, pc_3^{(i)}, ..., pc_m^{(i)})$ に最も類似している文書 (最大類似文書) の固有名詞成分の分布で近似する．類似度計算には，余弦類似度を用いる．

$$cos(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} {q_i}^2} \sqrt{\sum_{i=1}^{|V|} {d_i}^2}}$$

最後に，最大類似文書の分布から固有名詞成分をサンプリングする方法について説明する．

最大類似文書の主成分得点を正規化したものを累積化し，図 2 のように 0 ～ 1 までの固有名詞成分の累積分布を生成する．そこに 0 ～ 1 までの一様乱数を発生させ，その値に対応する固有名詞成分をサンプリングする．

以上が代表文書の選定法である．

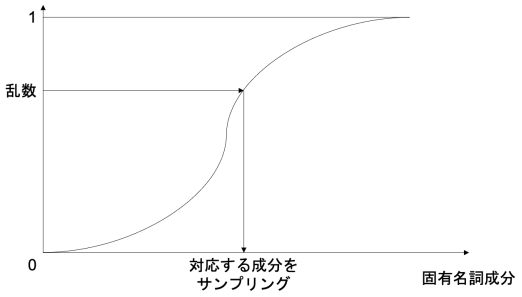

図 2　最大類似文書の累積分布

この手法では，最大類似文書の固有名詞成分分布からサンプリングを繰り返すことによって，カテゴリ全体の固有名詞成分分布を反映させている．訓練ベクトルの固有名詞成分分布が安定すると最大類似文書も安定するので，収束した訓練ベクトルの代表性は最大類似文書で判断する．収束した最大類似文書はカテゴリ全体の固有名詞成分の特徴を反映させているので，カテゴリ内を代表するトピックの文書であることが期待できる．

5. 実　　験

5.1 実験準備

本稿では CD-毎日新聞 2017 年版に採録されている 1 月 1 日から 1 月 14 日の 2 週間分のデータを抽出する．1 記事を 1 文書とする．固有名詞を抽出するため，形態素解析ソフト Mecab により形態素解析を行う．Mecab が扱う 69 の品詞体系のうち固有名詞を抽出し，抽出した固有名詞の出現頻度に従い文書をベクトル化する．固有名詞数が 5 語以上 100 語以下の文書を扱う．そうして得られた文書-固有名詞行列を主成分分析し，主成分得点で表された文書-固有名詞成分行列を生成する．単語のサンプリングにおいて主成分得点を正規化した値で累積化するので，マイナスの値を扱えない．よって，主成分得点-5 未満の固有名詞成分を含む文書を取り除き，全体に 5 を足す．文書数は 909 文書，固有名詞成分の数は累積寄与率 80 %までを扱い，383 成分である．

次に，同じ文書を扱って，Mecab で名詞を抽出し，名詞の中頻度語の出現頻度に従い文書をベクトル化する．そうして得られた文書-名詞行列は文書集合をカテゴリ分けするためのクラスタリングをする際に用いる．

5.2 評価方法

前述した通り代表文書を以下の 2 つの条件で定義する．1 つ目は，カテゴリ内の代表的な固有名詞成分が文書の特徴となる固有名詞成分であること，2 つ目は，文書の内容が時期特有の旬の話題であることである．よって，ギブスサンプリングで収束した最大類似文書の代表性を，以下のように評価する．

文書内で主成分得点が高い固有名詞成分の上位 10 成分を文書の主要成分とし，文書の特徴として扱う．また，クラスタ内

で文書の主要成分とされている文書数が多い固有名詞成分のうち，上位 10 成分をクラスタの主要成分とし，クラスタの特徴として扱う．代表文書の 1 つ目の条件については，最大類似文書の主要成分とクラスタの主要成分を比較し評価する．　2 つ目の条件については，文書の内容から 2017 年 1 月の旬の話題であるかを評価する．

5.3 結　　果

まず，毎日新聞記事を大まかなカテゴリに分けるためのクラスタリングをする．文書-名詞行列を TF*IDF で重みづけし，k-平均法を用いて 20 個のクラスタを生成する．結果を図 3 に示す．

このうち，クラスタ 3，クラスタ 6，クラスタ 10，クラスタ 19 においてギブスサンプリングによる代表文書の選定を行う．クラスタ 3 は文書数が多く，クラスタ内の文書の内容は大きくトランプ氏関連，経済，中国関連の 3 つに分かれている．クラスタ 6 は文書数が多く，クラスタ内の文書の内容にまとまりがない．クラスタ 10 は文書数は中程度で，クラスタ内の文書の内容は政治関連でまとまっている．クラスタ 19 は文書数が少なく，クラスタ内の文書の内容は事件や事故のカテゴリでまとまっている．

クラスタ	文書数	内容
1	14	戦争・軍事
2	47	政治/皇室/世界情勢
3	118	トランプ/経済/中国
4	50	スポーツ
5	44	韓国/外交
6	326	色々
7	12	社会問題(労働)
8	11	自動車
9	2	本
10	35	政治
11	27	世界情勢/外交
12	11	事件/テロ
13	88	芸術・文化、ライフ、コラムetc
14	8	将棋
15	11	事件
16	11	芸術・文化/競馬
17	4	スポーツ
18	51	災害
19	16	事件・事故
20	23	芸術・スポーツ

図 3　クラスタリング結果

クラスタ 3 の最大類似文書の遷移状況を図 4 と図 5 に示す．訓練ベクトルの要素数は，図 4 が 3830(383 成分× 10) で，図 5 が 11490(383 成分× 30) である．訓練ベクトルの要素数が 3830 の場合は収束していないが，訓練ベクトルの要素数が 11490 の場合は 14 文書の間で遷移するように収束している．これは，訓練ベクトルの要素数を増やすことによって文書の固有名詞成分の分布をより反映できるようになるためだと言える．訓練ベクトルの要素で分布を表現するので要素数を増やせばより正確に分布を表現でき，文書の収束を促すことができる．

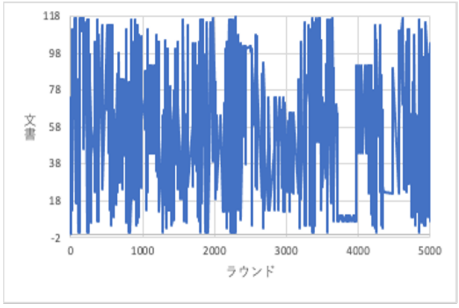

図 4　最大類似文書の遷移 (クラスタ 3, 要素数 3830)

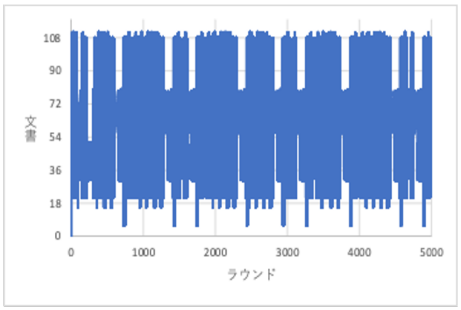

図 5　最大類似文書の遷移 (クラスタ 3, 要素数 11490)

図 6 はクラスタ 3 の主要成分であり，図 7 は図 5 で収束した文書の主要成分である．

クラスタ 3 の主要成分の文書数の割合 (クラスタ内で固有名詞成分 pc_i が主要成分とされている文書数をクラスタ内の総文書数で割った値) はあまり高くない．これは，クラスタの内容がトランプ氏関連，経済，中国関連と大きく 3 つに分かれていることが影響していると言える．収束文書の主要成分をクラスタの主要成分と比較すると，大部分の収束文書の主要成分にクラスタの主要成分の上位 3 成分が含まれており，内容はトランプ氏に関連する文書で共通していた．収束文書はクラスタの主要成分を反映しており，同じ話題の文書である．2017 年 1 月はトランプ氏がアメリカ大統領選に当選した直後で，旬の話題である．代表文書として相応しいと言える．

	1	2	3	4	4	6	7	7	9	9	9
成分	24	94	37	40	151	31	8	42	21	111	138
文書数	33	26	25	15	15	14	13	13	12	12	12
割合	0.28	0.22	0.212	0.127	0.127	0.119	0.11	0.11	0.102	0.102	0.102

図 6　クラスタ 3 の主要成分

	1	2	3	4	5	6	7	8	9	10
文書21	47	34	99	31	42	41	94	46	381	24
文書29	24	368	240	360	37	47	94	249	242	333
文書30	37	24	40	26	97	315	94	80	193	368
文書31	37	24	40	260	94	296	261	97	264	315
文書32	92	190	120	111	96	27	262	67	363	8
文書37	37	40	149	136	147	138	111	92	119	191
文書51	37	260	40	24	94	296	261	97	285	26
文書56	8	29	60	111	237	250	120	284	341	23
文書58	42	31	41	34	24	57	16	43	29	94
文書72	37	24	8	29	342	99	139	278	26	250
文書79	35	195	260	196	227	24	354	245	216	194
文書80	111	42	31	29	240	8	34	337	188	237
文書108	47	341	35	381	24	41	99	31	46	324
文書111	177	243	248	278	135	270	37	23	251	24

図 7　クラスタ 3 の収束文書の主要成分

図 8 はクラスタ 6 における最大類似文書の推移状況である．遷移回数が極端に少なく，一つの文書に収束している．

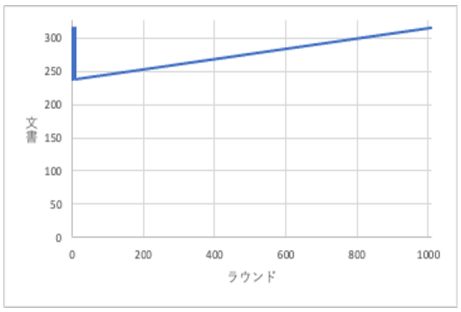

図 8　最大類似文書の遷移 (クラスタ 6, 要素数 11490)

図 9 はクラスタ 6 の主要成分であり，図 10 は収束文書の主要成分である．

クラスタ 6 の主要成分はクラスタ内における文書数の割合が極

	1	2	2	2	2	6	7	7	7	10	11
成分	360	13	16	344	383	14	284	294	305	23	371
文書数	23	21	21	21	21	20	19	19	19	18	18
割合	0.071	0.064	0.064	0.064	0.064	0.061	0.058	0.058	0.058	0.055	0.055

図 9　クラスタ 6 の主要成分

文書314	1	2	3	4	5	6	7	8	9	10
成分	308	300	292	210	186	340	204	281	275	262
主成分得点	12.17	9.706	9.561	9.111	8.705	8.619	8.385	8.353	8.333	8.325

図 10　クラスタ 6 の収束文書の主要成分

端に低く，文書の特徴とは言えない．これはクラスタ内に様々な内容の文書が混在していることが影響していると言える．収束文書の主要成分はクラスタ 5 の主要成分を全く反映していない．また，収束文書の主要成分の中でも主成分得点がひときわ大きな成分 308 は，クラスタ内で主要成分とされている文書数が 4 文書のみである．このことから，収束文書は代表文書とは言えず，ギブスサンプリングの条件付き確率分布の遷移の過程で特異な分布に収束していると考えられる．

図 11 はクラスタ 10 における最大類似文書の推移状況である．

約 1500 ラウンドの遷移の後，一つの文書に収束している．

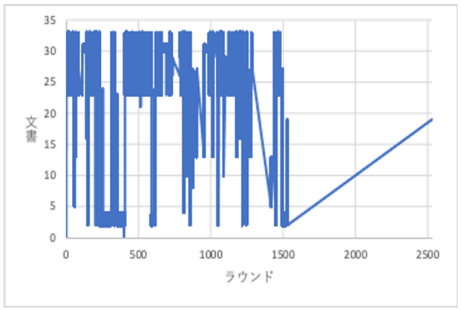

図 11　最大類似文書の遷移 (クラスタ 10, 要素数 3830)

図 12 はクラスタ 10 の主要成分であり，図 13 は収束文書と主な遷移文書の主要成分である．

クラスタ 10 の主要成分はクラスタ内における文書数の割合が

	1	1	3	4	5	5	7	7
成分	14	25	23	36	6	297	62	119
文書数	17	17	16	8	7	7	5	5
割合	0.486	0.486	0.457	0.229	0.2	0.2	0.143	0.143

図 12　クラスタ 10 の主要成分

	1	2	3	4	5	6	7	8	9	10
文書19	25	14	256	297	337	6	342	23	375	326
以下主な遷移文書										
文書2	25	297	256	14	183	6	293	234	294	141
文書4	25	14	256	297	183	293	234	6	230	337
文書5	25	271	14	378	156	23	6	197	184	33
文書10	25	14	119	359	282	23	303	138	374	280
文書13	38	25	119	128	130	14	162	23	9	175
文書23	14	25	23	36	13	297	310	294	62	235
文書24	14	25	36	23	344	13	330	342	370	30
文書27	25	14	23	156	30	6	9	20	182	54
文書30	36	14	25	173	23	319	311	208	196	96
文書31	36	14	25	173	23	319	311	196	208	13
文書33	14	25	23	36	13	297	310	62	235	294

図 13　クラスタ 10 の収束文書と主な遷移文書の主要成分

高く，特に上位 3 成分は 5 割に近い．クラスタ内の文書の内容が政治で統一されていることが反映していると言える．また，収束文書と主な遷移文書の主要成分と，クラスタ 10 の主要成分を比較すると，ほぼ全ての遷移文書でクラスタ 10 の主要成分の上位 3 成分が含まれている．収束文書はクラスタ 10 の主要成分を反映できている．主な遷移文書の内容は全て選挙で統一されており，収束文書は都知事選に関する内容であった．2016 年から 2017 年にかけて都知事選や衆院解散総選挙が話題になっており，旬の話題である．収束文書は代表文書に相応しいと言える．

図 14 はクラスタ 19 における最大類似文書の推移状況である．

4 つの文書の間で遷移するように収束している．

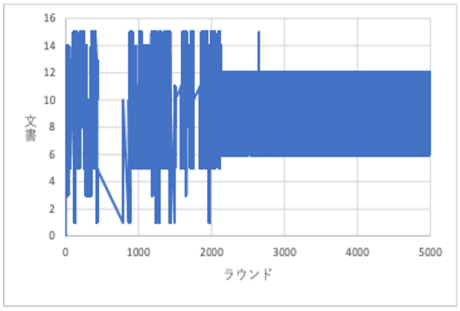

図 14　最大類似文書の遷移 (クラスタ 19, 要素数 3830)

図 15 はクラスタ 19 の主要成分であり，図 16 は収束文書の主要成分を表す．

クラスタ 19 の主要成分はクラスタ内における文書数の割合が高く，特に上位 2 成分は 3 割を超える．クラスタ内に類似した事件や事故が含まれていると考えられる．全ての収束文書の主

	1	2	3	3	3	3	3	3	9	9	9	9
成分	302	252	85	180	183	203	287	300	108	249	267	301
文書数	7	5	4	4	4	4	4	4	3	3	3	3
割合	0.438	0.313	0.25	0.25	0.25	0.25	0.25	0.25	0.188	0.188	0.188	0.188

図 15　クラスタ 19 の主要成分

	1	2	3	4	5	6	7	8	9	10
文書6	287	252	302	183	300	85	203	108	180	259
文書7	287	252	85	183	300	302	203	180	108	270
文書9	252	287	85	108	302	183	300	203	180	97
文書12	287	252	183	300	302	203	180	270	85	259

図 16　クラスタ 19 の収束文書の主要成分

要成分はクラスタ 19 の主要成分の上位 8 成分を含んでおり，クラスタの主要成分を正確に反映している．収束文書 4 文書の内容は同一事件である．事件内容は秋田県での死体遺棄事件で，何日にも渡ってこの事件を追っていることから旬の話題である．収束文書は代表文書として相応しいと言える．

6. 結　　論

本研究では，固有名詞成分とギブスサンプリングを用いた，カテゴリのトピックとなる代表文書の選定法を提案した．代表文書とは旬の話題でカテゴリの特徴を表したトピック文書であり，本研究では固有名詞成分の分布を文書の特徴として扱っている．

文書の内容にまとまりのあるクラスタでは，クラスタ内の固有名詞成分の分布の特徴を反映させた旬の話題である代表文書が選定できた．一方，文書の内容にまとまりのないクラスタでは，クラスタ内で固有名詞成分の分布が特異な文書に収束してしまった．これにはクラスタの主要成分の文書数の割合 (クラスタ内で固有名詞成分 pc_i が主要成分とされている文書数をクラスタ内の総文書数で割った値) が関連している．文書の内容にまとまりがあるクラスタには文書数の割合が高く特徴となる固有名詞成分があったが，文書の内容にまとまりのないクラスタには無かったことが原因である．

また，最大類似文書が収束しない場合，訓練ベクトルの要素数を増やすことによって収束を促すことができた．これは，訓練ベクトルの要素で文書の固有名詞成分分布を表現するので，要素数を増やせばより正確に分布を表現できるからであると考えられる．

文　　献

[1] 大森裕浩：”マルコフ連載モンテカルロ法の最近の展開” 日本統計学会誌 (2001)

[2] 樋山有理香, 三浦孝夫: ”n-TF*IDF による情報検索 ” 第 12 回データ工学と情報マネジメントに関するフォーラム (DEIM20), 福島県郡山市 ,2020

[3] 手塚太郎: ”しくみがわかるベイズ統計と機械学習”, 朝倉書店, 2019

[4] 須山敦志, 杉山将: ”ベイズ推論による機械学習入門” 講談社, 2018

[5] 高橋勝也, 三浦孝夫: ”k-平均法のための大域的クラスタ数決定.” 電子情報通信学会総合大会, 広島大学, 東広島, 2020

[6] 三浦大輝, 三浦孝夫: ”確率的 TF-IDF を用いた特徴語抽出と文書検索.” 2018 年情報処理学会全国大会, 2018, 東京, 6P-3

[7] 宮本定明: ”クラスター分析入門, ファジィクラスタリングの理論と応用” 森北出版株式会社, 1999