DEIM2021 H21-4

N-gram 処理を用いた代表文書の確率的生成

大野奈那子† 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: †nanako.ohno.2f@stu.hosei.ac.jp, ††miurat@hosei.ac.jp

あらまし 複雑かつ膨大な事象の文書データから代表文書を決めることで、その事象に対しての比較や研究を容易にすることを本研究の目的とする。本研究の手法として、MCMC(Markov Chain Monte Carlo) の 1 つであるギブスサンプリングを用い、文書データは 2 週間分の新聞記事を扱う。また、N-gram 処理 (FP-growth アルゴリズム) による語生成とギブスサンプリングによる語生成を組み合わせることによって、N-gram 性 (常識的な語のつながり) を持った代表文書を生成する。そして、選定した代表文書の記事内容は合理的か、N-gram 性を持っているかを評価する。また、ギブスサンプリングの問題点を解決するために、提案手法である N-gram 処理を用いたギブスサンプリングの有意性を示す。

キーワード 自然言語処理，代表文書，マルコフ遷移，Text Mining，Gibbs Sampling，N-gram

1. 前 書 き

近年、インターネットの普及により情報量の増加が見られている。この複雑かつ膨大な情報量を処理することは極めて難しい。そこで、テキストマイニングという大量の文書データから確率統計的な手法で有用な情報を取り出すテキストデータの分析方法がある。複雑かつ膨大な事象の文書データから代表文書を定めることで、その事象に対しての比較や研究を容易にすることを本研究の目的とする。また、本研究では新聞記事をデータとして代表文書を定めるため、ある記事の時期の特定や新聞社同士の比較など、新聞記事を用いたデータ分析を容易にすることも本研究の目的の 1 つとする。本論文では複数の文書をもとに、代表文書を生成する方法を提案する。

直接サンプリングが難しい確率分布の代わりにそれを近似するサンプル列を生成する方法で MCMC（Markov Chain Monte Carlo）の手法の 1 つにギブスサンプリングがある。[1] [2] このギブスサンプリングを用いて、複数ある文書データをもとに代表文書の確率的生成を行う。真の事後分布が得られるギブスサンプリングを用いることで、複数の文書から代表文書を選定できる。本研究では、複数の文書の対象を毎日新聞 2 週間分の記事にして代表文書を選定する。新聞記事は比較的、文体が統一されているためテキストマイニングに適している。新聞記事の代表文書により、各新聞社の比較や、ある特定の時期の代表記事が得られる。しかし、ギブスサンプリングを用いて選定された代表文書の性質は不明であり、複数の文書の要約なのか、特徴的な文書なのか利用方法が自明でない。

本論文では、以下のように構成される。第 2 節では、代表文書の定義、第 3 節では、簡単な例を用いた解法の説明、第 4 節では、解法の数式化、第 5 節では、N-gram 表の説明、第 6 節では、実験手順とその評価、実験結果、考察、第 7 節で結論を述べる。

2. 代 表 文 書

文書集合が全体として共通の特性を有すると仮定する。たとえば、”毎日新聞の記事”や”シェークスピアの小説”を考えればよい。このとき、次の 4 つの性質を考える。

(1) 内容がよくみられるトピックである。
(2) 新聞記事に多く出現する単語が含まれている。
(3) 常識的な語のつながりを持った語列を重視している。
(4) 確率的に代表文書を生成できる。

本稿では、この 4 つの条件を満たしているものを”代表文書”と定義する。文書集合に馴染まない内容であったり、（ギブスサンプリングなどの) 確率的生成が機能しない状況であったりした場合は、対象とする文書ではないとする。

3. ギブスサンプリング

複数の新聞記事の中からギブスサンプリングを用いて、代表文書を選定できることの意味を論じよう。

新聞記事を文書ベクトル化し、この文書ベクトルを文書と呼ぶ。ギブスサンプリングを行うには、初期値を設定する必要がある。単語数 30 としてランダムに単語を発生させ、出現頻度をそれぞれ 1 にしたものを初期値とする。そしてギブスサンプリングは、事前にサンプルされたものを確率分布として、確率変数を 1 つずつサンプルする方法であるため、事前にサンプルされたものから確率分布を条件付けする必要がある。本研究では、確率変数をサンプルするための手法として、余弦類似度を用いる。余弦類似度とは、文書同士を比較する際に用いられる類似度計算手法で、1 に近いほど内容が類似している。この余弦類似度を用いて、（ギブスサンプリングの）事前確率分布とすべての記事の類似度を計算する。類似度最大の文書から語を生成する。

ギブスサンプリングを用いた代表文書の生成方法を示す。語の生成にはモンテカルロ法を用いる。あらかじめ計算された新聞記事の単語の出現確率を用いて、累積確率分布を求め、これ

に従う乱数を生成することにより特定の記事に出現する単語だけを生成することが可能となる。このようにして、先頭の単語を置き換える。同じように繰り返しサンプリングを行い、これを続けると代表文書に収束する。

しかし、この代表文書の生成方法には問題点がある。1 語ずつ置き換えるため語同士の依存性 (N-gram 性) を考慮していない点である。N-gram 性とは、2gram 以上で意味をなすものを指す。例えば、2gram 性を持っていると言える文書の例として「ドナルド・トランプ」が挙げられる。「ドナルド」という 1gram だけでは「ドナルド・ダッグ」なのか「ドナルド・トランプ」なのか判定ができない。このような 2gram 性をマルコフ性と考える。マルコフ性とは、確率論における確率過程の特性の 1 つで、その過程の将来状態の条件付き確率分布が、現在状態のみに依存する特性を持つことをいう。例えば、「ドナルド」に続いて「トランプ」が生起する確率が定まれば、「ドナルド・トランプ」という語列はマルコフ性を持つということになる。また、3gram 性を持つ文書の例として「へそで茶を沸かす」という慣用句が挙げられる。「へそ・茶・沸かす」という 3gram になって、慣用句としての意味が初めて成立する。本研究では、2gram 以上で意味をなすもの（N-gram 性）を考慮するための N-gram 表を作成する。ギブスサンプリングと N-gram 表の双方を用いて単語を生成することで、常識的な語のつながり（N-gram 性）を持った代表文書の生成が期待できる。

4. N-gram 処理を用いたギブスサンプリング

本章では、3 章で提案したギブスサンプリングを用いた代表文書の生成方法を数式で説明する。初期値を 1 にした場合のギブスサンプリングを考える。

$$d_0 = (w_1^{(0)}, w_2^{(0)}, w_3^{(0)}, w_4^{(0)})$$

ここで、d_0 の 1 番目の単語 $(w_1^{(0)})$ を置き換えたい。数式①を用いて単語を生成する。

$$w_1^{(1)} \sim p(w_1 \mid w_2^{(0)}, w_3^{(0)}, w_4^{(0)}) \cdots ①$$

この $p(w_1 \mid w_2^{(0)}, w_3^{(0)}, w_4^{(0)})$ は事前にサンプルされている $(w_2^{(0)}, w_3^{(0)}, w_4^{(0)})$ と余弦類似度最大の文書から w_1 を生成するという意味である。同じように語を置き換えると下記のように全ての語が入れ替わる。

$$\begin{aligned}
&(w_1^{(1)}, w_2^{(0)}, w_3^{(0)}, w_4^{(0)}) = d_2 \\
&w_2^{(1)} \sim p(w_2 \mid w_1^{(1)}, w_3^{(0)}, w_4^{(0)}) \\
&(w_1^{(1)}, w_2^{(1)}, w_3^{(0)}, w_4^{(0)}) = d_3 \\
&w_3^{(1)} \sim p(w_3 \mid w_2^{(1)}, w_3^{(1)}, w_4^{(0)}) \\
&(w_1^{(1)}, w_2^{(1)}, w_3^{(1)}, w_4^{(0)}) = d_4 \\
&w_4^{(1)} \sim p(w_4 \mid w_1^{(1)}, w_2^{(1)}, w_3^{(1)}) \\
&(w_1^{(1)}, w_2^{(1)}, w_3^{(1)}, w_4^{(1)}) = d_5
\end{aligned}$$

この過程を 1 ラウンドとし、これを収束するまで繰り返す。また、N-gram 性を反映させた語生成を行うため、以下の手法で代表文書を生成する。

同じく初期値を d_0 にした場合の N-gram 処理を用いたギブスサンプリングを考える。1 番目の単語は、同じように通常のギブスサンプリングから語を生成するため、①を用いて、w_1 を生成する。次に 2 番目の単語を置き換える時から、確率 ε の割合で、N-gram 処理を行う。ギブスサンプリングの処理に、確率 ε の割合で N-gram 処理を加えることによって、幅広い知識を獲得することをここでは、 ε -greedy とする。本研究では、 ε を 10%に設定して N-gram 処理を用いたギブスサンプリングを行う。N-gram 処理とはギブスサンプリングから語生成を行うのではなく、N-gram 表から語生成を行う処理のことであり、N-gram 表について詳しくは次章で言及する。ここで、N-gram 表について、簡潔に説明しておくと、同時に出現する確率が閾値以上である N-gram を収集したものである。例えば、”ドナルド”と”トランプ”が同時に出現する確率が、ある閾値以上であれば、N-gram 性（ここでは 2gram 性）があるとみなされ、N-gram 表に追加される。このようにすべての同時出現確率を求めて、N-gram 表を作成していく。このように作成した N-gram 表を用いて語生成を行う。前述したように、2 番目の単語からギブスサンプリングで語生成を行うか、N-gram 表から語生成を行うか条件分岐する。N-gram 処理で語生成を行うと分岐された場合は、ランダムに N を定める。そして、基本的に 10%の割合で N-gram 表を用いて語生成を行うのだが、N-gram 性を持たない単語もあるため、N-gram 性を持たない場合は、ギブスサンプリングを行うという条件とする。10%の割合で N-gram 処理で語生成を行う例は以下となる。まず、1 番目の単語はギブスサンプリングで語生成を行うため、d_2 の状態から考える。ここで置き換えたい単語は w_2 であるため、それ以外の語列 $(w_1^{(1)}, w_3^{(0)}, w_4^{(0)})$ をみる。本研究に用いる N-gram 表は 6gram まで存在するため、2 から 6 までの数字で乱数発生させる。例えば N=3 に決まった場合、その文書に 3gram になりうる 2gram が存在するかを調べる。この場合、$(w_1^{(1)}, w_3^{(0)})$,$(w_1^{(1)}, w_4^{(0)})$,$(w_3^{(1)}, w_4^{(0)})$ の 3 つの 2gram が考えられる。もし、N-gram 表に $(w_2 \mid w_1^{(1)}, w_4^{(0)})$ となるような 3gram が存在すれば、以下のように語生成を行う。

$$w_2^{(1)} \sim p(w_2 \mid w_1^{(1)}, w_4^{(0)})$$

もし、3gram になりうる 2gram が複数存在していた場合は、ランダムに選んだ 2gram から語生成を行うものとする。このように N-gram 処理を用いたギブスサンプリングを行うことで、ギブスサンプリングの問題点であった N-gram 性の破壊を抑えることができ、常識的な語列を配慮した語生成ができる。

5. N-gram 表

N-gram 処理に用いられる N-gram 表は FP-growth アルゴリズムで作成する。FP-growth とは、頻出パターン抽出手法の 1 つでトランザクションデータベースを圧縮したデータ構造である FP-Tree から、頻出アイテムセットを抽出する手法である。[3] [4] [5] [6] この FP-growth アルゴリズムを用いて、新聞記事を元データとして N-gram 表を作成する。本研究では、出現回数を 50 回以上、信頼度を 0.5 以上とする。

表 1　N-gram 表（一部）

2gram	3gram	4gram
容疑，事件	優勝，試合，大会	トランプ，大統領，就任，次期
原発，事故	会見，発表，記者	大統領，日本，経済，トランプ
ラグビー，大会	五輪，東京，日本	中国，大統領，米，米国
団体，日本	安倍，晋，首相	政権，米国，経済，トランプ

表 1 が本研究に用いる N-gram 表の一部である。

6. 実　　験

6.1 実験手順

本稿では、データとして「CD-毎日新聞 2017」の 1 月 1 日〜1 月 14 日の 2 週間分の記事を使用する。総文書数は 2443 文書で、その中から総述べ語数が 5 語未満の文書を除いたもの 2302 文書を使用する。また Mecab で形態素解析を行い、名詞のみを抽出した 547 語である。

6.2 代表文書とその評価

ギブスサンプリングで生成された代表文書の記事内容と出現単語、類似文書の移り変わり、および代表文書の N-gram の割合で評価する。代表文書の記事内容に関しては、代表的といえる内容かどうか読んで評価を行う。類似文書の移り変わりが合理的であれば、ギブスサンプリングが有効に機能しているとする。最終的に生成された代表文書の N-gram の割合を調べ、常識的な語のつながりをもつ代表文書が生成できたか評価する。通常のギブスサンプリングで生成された代表文書との比較も行う。

6.3 実験結果

通常のギブスサンプリングで生成された代表文書と N-gram 処理を用いたギブスサンプリングで生成された代表文書の記事内容をそれぞれ以下に示す。

安倍晋三首相＝似顔絵＝は２日、神奈川県茅ケ崎市のゴルフ場で、経団連の榊原定征会長ら財界人と今年初めてのゴルフを楽しんだ。記者団に「幸先の良いゴルフになってますよ」と語った。また、早期の衆院解散・総選挙の可能性について問われると、笑いながら「ない、ない」と否定した。　首相は１日は、夫人の昭恵さん、岸信夫副外相ら親族と映画「海賊とよばれた男」を鑑賞。３日までは休み、４日に伊勢神宮を参拝して年頭の記者会見に臨む。【真野敏幸】

図 1　代表文書の記事内容（通常）

【ロサンゼルス長野宏美】米女優メリル・ストリープさんが映画賞のゴールデン・グローブ賞の授賞式でトランプ次期大統領を批判したことに対し、トランプ氏の支持者らが反発している。ソーシャルメディアでは「ボイコット・オスカー」や「ボイコット・ハリウッド」などのハッシュタグができ、２月に開かれるアカデミー賞の授賞式を視聴しないよう呼びかけている。　「現実離れした大富豪が賞を与え合う（アカデミー賞）授賞式なんて見ない」「（次期）米大統領を尊敬しないエリートに対して立ち上がる」。トランプ氏の支持者らはツイッターでこう書き込んでいる。大統領選で広がった「反エリート」の波がハリウッドを襲い、米国の分断を浮き彫りにしているようだ。　多様性やメディアの役割の重要性を訴えたストリープさんの演説は称賛の声が起きた。その一方で、「反トランプ」を公言していた人からは、「警告」も出ている。　「メリル・ストリープの演説こそ、トランプが勝った理由だ」。２００８年の大統領選で共和党候補だったマケイン上院議員の娘で、反トランプの立場から今回は無党派に投票したというメーガン・マケインさんはツイッターに投稿。「ハリウッドの人たちが（トランプ氏が勝利したのは）なぜかを認識しなければ、彼の再選を助けることになる」と主張し、エリートが正論を吐くだけでは大衆の心に響かず、トランプ氏の支持者に目を向けるよう促した。

図 2　代表文書の記事内容（N-gram 処理）

図 1 から安倍首相のある日の行動について書かれている記事であることがわかる。”首相”や”衆院”、”総選挙”など新聞記事によく見られる政治関連の単語が含まれている。図 2 からアメリカのトランプ政治について書かれている記事であることがわかる。この記事も図 1 と同じく”大統領”や”トランプ”、”支持者”などの政治関連の単語が多く含まれている。次に、通常のギブスサンプリングでの類似文書の移り変わりと N-gram 処理を用いたギブスサンプリングの類似文書の移り変わりを以下に示す。

類似文書	テーマ	最大類似度
1681	政治（台北・トランプ）	0.2081546
267	政治（トランプ）	0.2188314
1990	政治（トランプ）	0.65098252
112	政治（安倍ゴルフ）	0.88982423

図 3　類似文書の移り変わり（通常）

類似文書	テーマ	最大類似度
1681	政治（トランプ）	0.208154596
2265	政治（トランプ）	0.297466347
859	政治（トランプ）	0.468847985
223	政治（トランプ）	0.480537231
859	政治（トランプ）	0.512882017
1984	政治（トランプ）	0.529957733
223	政治（トランプ）	0.545798597
1778	政治（トランプ）	0.59830149
115	政治（トランプ）	0.597824218
1778	政治（トランプ）	0.786011774
2021	政治（トランプ）	0.806559133
1778	政治（トランプ）	0.805939006
2021	政治（トランプ）	0.789731777
一部省略		
664	政治（トランプ）	0.718144217
1778	政治（トランプ）	0.746567131
979	政治（トランプ）	0.710981368
1778	政治（トランプ）	0.684655781
223	政治（トランプ）	0.640686221
1322	政治（トランプ）	0.646442316
1985	政治（トランプ）	0.649238107
115	政治（トランプ）	0.649123561
1985	政治（トランプ）	0.639281512
115	政治（トランプ）	0.624736846
1985	政治（トランプ）	0.631613941
115	政治（トランプ）	0.642885192
859	政治（トランプ）	0.631861913
223	政治（トランプ）	0.642038738
1985	政治（トランプ）	0.641805963
223	政治（トランプ）	0.695865446
1778	政治（トランプ）	0.840024172

図 4　類似文書の移り変わり（N-gram 処理）

そして、通常のギブスサンプリングと N-gram 処理を用いたギブスサンプリングによって生成された代表文書の N-gram 性

を持つ単語数を表 2 に示す。表 2 から、N-gram 処理を用いたギブスサンプリングで生成された代表文書の N-gram の割合は、通常のギブスサンプリングで生成された代表文書の N-gram の割合の 4 倍であることがわかる。また、通常のギブスサンプリングでは 2gram と 3gram しか含まない結果に対して、N-gram 処理を用いたギブスサンプリングでは 2gram から 6gram まで含んでいることがわかる。

表 2　N-gram 表（一部）

	ギブスサンプリング（件）	N-gram 処理（件）
2gram	6	3
3gram	1	5
4gram	0	11
5gram	0	10
6gram	0	1

6.4 考　　察

図 1、図 2 を見ると、通常のギブスサンプリングで選定した代表文書も N-gram 処理を用いたギブスサンプリングで選定した代表文書もどちらも新聞記事によく見られる政治記事であることがわかる。また、政治記事によく見られる政治関連の単語を多く含んでいる。それぞれの特徴として、文書サイズに差があることが挙げられる。N-gram 処理を用いたギブスサンプリングの代表文書は平均的な文書サイズであることに対して、通常のギブスサンプリングの代表文書は、比較的小さいサイズの文書であることがわかる。記事内容や出現単語については、どちらも新聞記事にふさわしい内容であるため、代表的であると考える。

図 3、図 4 を見ると、通常のギブスサンプリングの方が、文書の移り変わりが少ないことがわかる。これは、N-gram 処理を用いたギブスサンプリングは 10%の割合で、N-gram 表から語生成を行うため、収束するまでにより時間がかかったと考える。また、通常のギブスサンプリングと N-gram 処理を用いたギブスサンプリングの類似文書はどちらも政治記事に一貫していることがわかる。しかし、さらに詳しく記事内容を見てみると、通常のギブスサンプリングでは、台北の政治について書かれた記事やトランプ政治、安倍政治など内容に一貫性が欠けていることがわかる。一方、N-gram 処理を用いたギブスサンプリングでは、トランプ政治について書かれた記事に一貫していることがわかる。この類似文書の移り変わりの結果から、N-gram 処理を用いたギブスサンプリングの方が、代表文書の確率的生成の精度が高いと言える。

そして、表 2 から代表文書の N-gram の割合は、N-gram 処理を用いたギブスサンプリングの方が通常のギブスサンプリングに比べて 4 倍件数が多くなっていることがわかる。この結果から、N-gram 処理を用いたギブスサンプリングの方が、N-gram 性（常識的な語のつながり）を重視した代表文書が生成できたと言える。

表 3 と代表文書の定義から、N-gram 処理を用いたギブスサンプリングの方が通常のギブスサンプリングより有意な結果を得られると言える。

表 3　評 価 要 約

	ギブスサンプリング	N-gram 処理
記事内容	政治記事（安倍）	政治記事（トランプ）
出現単語	政治関連	政治関連
移り変わり	一貫性○	一貫性◎
N-gram（比率）	1	4

7. 結　　論

本研究において、2 週間分の毎日新聞データをもとに、N-gram 処理を用いたギブスサンプリングによる代表文書の生成ができた。しかし、初期値依存性や収束の判断が難しいという問題が挙げられる。本研究では、初期値をランダムに設定、バーンイン処理を 1000 回に設定したが、より多くのバーンイン期間を設けることでより精度の高い代表文書の生成が可能になると思われる。また、データの拡大を行うことによって、N-gram 表の精度も上がり、より常識的な語列を重視した代表文書の生成が可能になると考える。

文　　献

[1] 須山敦志，杉山将: ”ベイズ推論による機械学習入門 ”, 2017

[2] 手塚太郎: ”しくみがわかるベイズ統計と機械学習 ”, 2019

[3] J. Han, J. Pei, and Y. Yin: ”Mining frequent patterns without candidate generation,” In Proc. of the ACM SIGMOD Conf. on Management of Data (2000)

[4] Sengly Heng, and Susumu Shibusawa: ”FP-Tree 分割による頻出アイテムセットの抽出” 第 19 回データ工学ワークショップ (DEWS 2008)

[5] 岩橋永悟，平手勇宇，山名早人: ”PC クラスタ上における頻出飽和パターン抽出並列化手法の提案”, 電子情報通信学会第 16 回データ工学ワークショップ (DEWS 2005)

[6] 平手勇宇，岩橋永悟，山名早人: ”TF2 P-growth:閾値設定を必要としない頻出アイテムセット抽出アルゴリズム” 情報処理学会論文誌：データベース (2005)