文書トピックの局面追跡

Tracking Aspects in Document Topics

井上　眞乙† 　三浦　孝夫††

† 法政大学大学院理工学研究科システム理工学専攻　〒 184–8584 東京都小金井市梶野町 3–7–2
†† 法政大学大学院理工学研究科システム理工学専攻　〒 184–8584 東京都小金井市梶野町 3–7–2
E-mail: †maoto.inoue.7h@stu.hosei.ac.jp, ††miurat@hosei.ac.jp

あらまし　本稿では文書トピックの局面追跡を論じる. 大規模な量のデータの中には断片的な情報が多く存在しているが, これらから人手でトピック内の状況の変化を把握する事は困難である. 本研究では隠れマルコフモデル (HMM) を用いる. 主たるアイデアは,HMM による状態遷移の過程がトピック内のストーリーの構造と捉えると仮定し, 尤度により類似する記事を選定・追跡することにある. フィルタリングのため,KL 情報量を用いる.

キーワード　隠れマルコフモデル, HMM, トピック追跡

1. 前　書　き

現在, 膨大な量の情報が日々増加している. 中には断片的な情報も多く存在し, 時間の経過とともにその内容が変化していく. トピック追跡はこのような断片的な情報から別の情報への内容の変遷を捉えることを目的としている. しかし, 大規模な量のため組み合わせ作業を人手で行う事は容易ではなく, 瞬時に流れ去っていく. 正しい追跡を実現するためには当然, 効率的かつ正確で有用な技術が必要になる.

トピックに関する代表的なアプローチの 1 つとして,Topic detection and Tracking (TDT) がある [4].TDT はある時点で発生する固有の事象を検出し, また事象の進展を追跡するための技術の研究を目的としている. これにはラジオのニュース情報や新聞記事など, 複数の情報源が想定されている.TDT における Event Tracking タスクでは, 特定のトピックに対するストーリーを特定する.Allan らは適合性フィードバックなど二つのアプローチを適用し, 情報フィルタリングに類似した手法を提案している [1].Carbonell らは決定木（dtree）と k 近傍法（kNN）を用いた追跡を行っている [9]. しかしこれらは事前にある程度の学習データを必要としている.

本研究では時系列データを扱う. 単純な構造を持つデータとは異なり時間の流れによって事象が変化するデータ, すなわち多くの場合は音声や新聞記事のような系列データである. これに対し HMM, 最大エントロピーモデル（MEMM）, 条件付確率場（CRF）などの確率論に基づく高度な手法がいくつか存在している.HMM は少量の学習データがあればより HMM がうまく機能する助けになる.MEMM や CRF は, 系列データに対し強力なモデルであるが事前に学習データが必要である. 学習データの取得は多くの時間とコストが掛かるため容易でないことは明らかであり, 大量の学習データを取得することはほぼ出来ない.

本研究では, 隠れマルコフモデル (HMM) を用いた尤度によるトピックの局面追跡を行う手法を提案する. さらに,K-Viterbi アルゴリズム（及び後ろ向きアルゴリズム）による尤度の計算方法を提案する. これによりモデルの多重化が可能となり, 上位 K 個の尤度を持つ経路を推定することが可能となる. 本研究の貢献は以下の 3 つである.

(1)　文書トピックの局面を追跡し, 正確に内容の変遷を捉えるアルゴリズムを構築する.

(2)　学習データを必要とせず, 特定のトピックに対する内容を追跡する.

(3)　トピックの構造を抽出し, 全体像の把握を行う.

本論文は次のように構成されている.2 章ではトピックの局面追跡について,3 章では HMM について述べる.4 章では HMM によるトピックの局面追跡のアプローチを提案し,5 章では実験結果を示し,6 章では結論を述べる.

2. トピックの局面追跡

トピックとは, ある一連の事件を表す特定の話題のことである. 例えば「市川・女性刺殺事件」と「山口・周南放火殺人事件」の殺人事件に関する 2 つの記事があるとする. どちらも事件発生から容疑者起訴までの同様の一連の流れ（ストーリー）を持っているが, 凶器や被害者・容疑者の名前, 犯行場所や犯行手口等が異なる. つまり, これらを同様の「殺人事件」として捉えるのではなく,「市川・女性刺殺事件」と「山口・周南放火殺人事件」とする別々の事件を表す特定の話題として捉える方が自然である.

本研究ではトピックの局面追跡を行う. トピックの局面とは, トピックのストーリーを構成する個々の事象（イベント）のことである. トピック「千葉・市川女性刺殺事件」の記事では局面は「事件発生」,「犯人逮捕」,「犯人の供述」,「被害者の人柄」,「犯人の供述」,「容疑者起訴」である. この局面を追跡することで, トピックのストーリーを捉え, 発生から起訴までの全体像を分析することが可能となる. 北田らは,PLDA を用いてトピックを 1 日ごとに並列に抽出し, 余弦類似度に基づいて関連付けることによりトピック系列を作成, 可視化してトピックの

追跡を行う手法を提案している [11].

しかし, トピックの局面追跡には主に 2 つの問題がある. 第 1 にストーリーの変遷を捉えることが困難である. ストーリーの決まった構造が無いため学習データを用意する方法が無く, また同じジャンルのトピックであればストーリーの変遷が似ているため特定のトピックに対するストーリーの捕捉が難しい. 第 2 にトピック内で, ある事象がバースト状態（急激に記事数が増加する）が生じることがある. 急激な変化に対する対応が難しく, 同じトピックの記事の内容でありながら「違うトピックの記事である」と間違った追跡を行う場合がある.

本研究ではこの問題に対し,HMM による確率論的アプローチをとる. 状態の遷移をストーリーの構成として捉えマルコフモデルを構築し, ストーリーの構造の可視化を可能とする. さらに, 尤度計算でトピックの局面追跡を行う. 尤度上位 K 個の経路を同時に推定する K-Viterbi アルゴリズムと, 対になる後ろ向き K-Viterbi アルゴリズムを提案しモデルの多重化を図る.

例としてトピック「千葉・市川女性刺殺事件」の局面追跡を行う. 図 1 に「千葉・市川女性刺殺事件」の記事の一部を, 図 2 に「山口・周南放火殺人事件」の記事の一部を示す. これらはそれぞれ時系列順になっており,1 行が 1 記事に対応している. 語は固有名詞と動詞のみを抽出している.

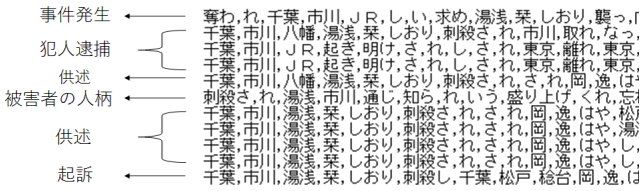

図 1　千葉・市川女性刺殺事件の記事

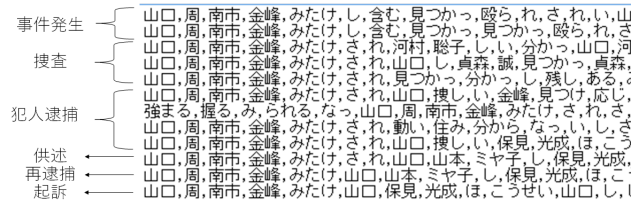

図 2　山口・周南放火殺人事件の記事

3.　隠れマルコフモデル

3.1　マルコフモデル

確率過程とは, 確率変数が一組の状態に関連付けられ, 時間の経過と共にモデルがランダムに変化する数学的な手法を意味し, 確率的有限オートマトン（出力観測値付）と考えられる. このアプローチは, 様相がランダムに変化する現象及びシステムの数学的モデルに広く使用されている. 一般的な例として, 分類問題や音声認識などが挙げられる. マルコフモデル（MM）は未来の状態が現在の状態のみに依存し, それ以前に発生した状態/出来事の影響は受けないという仮定を伴う確率論の枠組みの 1 つであり, マルコフ性と呼ばれる. 一般的にこの仮定は, 単純で効率的な方法で予測モデリングや確率的推測を容易にし, 扱いにくいシステムへの推論や計算を可能にする [3]. しかし, マルコフモデルには 2 つの深刻な欠点がある. どの状態にいるのかを推定する手段が無いことと, 学習データを構築する方法が不明な事である. 正解が与えられていないため学習データによってモデルを構築し, 状態遷移と出力観測値を求める必要がある. しかし, 作成者の主観に依存してしまい客観性に欠けてしまう. 真の問題は, 正確に, かつ効率的にモデルを構築する方法である.

3.2　隠れマルコフモデル

隠れマルコフモデル (HMM) とは, 状態が部分的にしか観測できないマルコフモデルの一種である. 出力観測値はシステムの状態に関連付けられるが, 状態を正確に決定するには不十分である. では, どのような潜在意味が, 最も適した解釈に対する観測値と関連した状態であるのか？HMM は 3 つの問題に対する解決策を持つ.観測系列 $O = o_1 ... o_N$ と仮定する. 第 1 に, 観測系列 O の確率を得る事が可能である. 実際には, マルコフモデルを仮定すると,O を生成する全ての経路を探索し, その確率の合計をとる. 第 2 に, 観測系列 O を生成する可能性が最も高い経路を取得する事が可能である.Viterbi アルゴリズムにより, 動的計画法を用いて効率的に最尤経路を推定する. また, 前向きアルゴリズムは観測系列の確率を計算する.

第 3 に, マルコフモデルの推定が可能なことである. つまり状態遷移確率 $P(s_j|s_i)$, 状態 s_i における観測値 o_k の確率 $P(o_k|s_i)$ 及び開始時の初期状態確率 $P(s_i)$ を推定する事が可能である. 基本的に HMM を推定するために EM アルゴリズムを適用する. 始めから EM アルゴリズムによる調整過程を繰り返すことにより, データの尤度を最大化していく. 事前に学習データを必要としない. 隠れマルコフモデルにはいくつか知られたアルゴリズムが存在し, 特に Baum Welch（BW）アルゴリズムは, (学習されていない）データ集合から効率的にマルコフモデルを推定する.

代表的な応用事例に,Viterbi アルゴリズムを利用し形態素解析を行う方法がある. つまり, 品詞に基づくマルコフモデルを構築し各経路の尤度（Viterbi アルゴリズムによる）を得て, 最大尤度となる経路を決定する. 例えば, 品詞のマルコフモデルと”Time flies like an arrow” の英文が与えられたとする. この英文の翻訳は単語”flies” と”like” の品詞が「名詞または動詞」,「動詞または前置詞」であるかによって「時蝿は矢を好む」または「光陰矢の如し」というように大きく差が出来てしまう. まず, 出力観測値として考慮したこの英文を区切り, 名詞, 動詞, 冠詞, 形容詞, 前置詞の状態遷移を推測する. 次に, 最大の尤度である状態の経路を決定する. この場合, 最尤経路は名詞, 動詞, 前置詞, 冠詞, 名詞であるため最適な翻訳は「光陰矢の如し」であると推測できる [10].

本研究では HMM を用いてモデルを構築し尤度の計算を行い, 状態遷移の過程をストーリーの構成と捉えてトピックの局面追跡を行う.

例として図 1,8 の記事を用いる. 尤度の計算を行う前にフィルタリングを行い, 得られた記事によりモデルを構築する. フィルタリングについては 4 章で述べる. 得られた記事の一部と用い

てその記事により構築したモデルを, 図 3 に示す.7 件の記事から HMM の Baum Welch アルゴリズムにより、モデルを推定する. 例えばこれらの記事で構築したモデルは初期状態確率が 0.9943... で状態 0 から始まるモデルが構築されている (実際の実験では状態名を各状態からの出力記号に基づいて決定する).

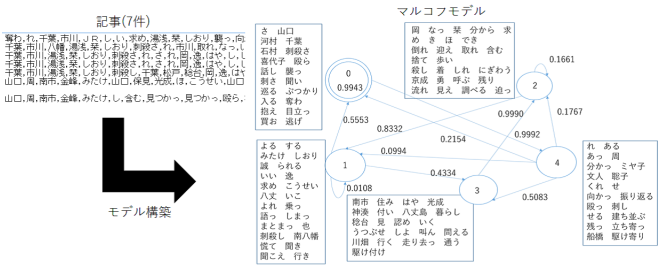

図 3　記事の一部と構築されたマルコフモデル

4.　HMM によるトピックの局面追跡

本章では文書とする新聞記事に対し,HMM を用いるトピックの局面追跡の新しいアプローチを提案する. 本研究は, ある記事が特定のトピックにおける記事の後に追跡して良いかを 1 つのモデルで判断するアプローチのため, クラスごとにモデルを構築し最尤原理 (MLP) により分類する分類問題とは異なることに注意したい.

Yang 氏らは kNN を適用し, 余弦類似度に基づいて追跡を行う手法を述べている [4][9]. 各文書をストーリーとみなし入力ストーリーをベクトル化し, 余弦類似度に基づいて求められた再近傍の k 個の各トピックの学習ストーリーの類似度の差が閾値以上であればターゲットとしているトピックのストーリーとみなす. 共に提案されている決定木による手法よりも結果がやや優れており, 修正版では誤検出率も減少出来ると述べている. しかし, 具体的な再現率や適合率までは述べていなく, 追跡した結果を明示して視覚化していない.

ここでは以下の 2 つの仮定を置き, 尤度によるトピックの局面追跡を行う手法を提案する.

(1)　トピックには統一した構造がある.

(2)　各トピック内で時系列順に並べられた文書には一定の尤度がある.

さらに,HMM の尤度計算において K-Viterbi アルゴリズム（及び後ろ向き K-Viterbi アルゴリズム）を提案する.K-Viterbi アルゴリズムとは, 尤度の大きい順に K 個の経路を推定するアルゴリズムである. これにより複数の経路を同時に推定することでモデルの多重化を図ることが可能となる.

（Step1）初期状態 S として可能な状態を P とする. この時のコスト $C_0(S) = [d_1(S, P_1), ..., d_k(S, P_i)]$,$(i = 1, ..., N)$

$m = 1, ..., M$ において,

（Step2）次の状態 Q と, その状態に入力する状態 P_i に対してコストを計算する.$C(P_i) = [k$ 個$] = [d_1(P_i, Q), .., d_k(P_i, Q)]$

（Step3）各 Q において,$MAX_k[C_{m-1}(P_i) + d(P_i, Q)]$ を計算し, 大きい順に K 個の尤度をリスト $C_m(Q)$ に格納する.

（Step4）Step2〜Step3 を繰り返す.

i とは状態数のことであり,M は観測値のサイズ,k は尤度の上位 k 個を示す.$d_1(P_i, Q)$ は状態 P_i から状態 Q への状態遷移確率×状態　Q における出力語 m の出力記号確率である. 上記の過程を繰り返すことで, 尤度上位 K 個の経路を得ることが可能となる.

例えば図 3 のモデルにより, 記事の尤度上位 2 個の経路を得るとする. 記事の中身は” 山口, 周, 南市, 金峰, みたけ, し, 含む, 見つかっ,..., いう” となっている. 最初の語” 山口” はモデルの’ 状態 0 のみから出力記号確率 0.0746... で出力されている. この時の初期状態 S のコスト $C_0(S)$ は, 初期状態確率と出力記号確率を掛け合わせた 0.074... となる. さらに語” 山口” から次の語” 周” へは状態 0 から状態 4 のみの遷移になる. これに対するコスト $C(P_i)$ は先程求めたコスト 0.074... を状態遷移確率と出力記号確率と共に掛ける. 得られた値を大きい順に 2 個リストに格納すると, 最大で 0.0022... となり, 経路は状態 0 から状態 4 へのみの遷移である. これを最後の語まで計算していく.

後ろ向き K-Viterbi アルゴリズムは,K-Viterbi アルゴリズムを後ろ向きに行うアルゴリズムである. 本研究ではモデルの状態遷移をストーリーの遷移とみなすが, 後ろの内容の記事をモデルの前から Viterbi アルゴリズムで行うと正しく尤度が計算できないのではという恐れがある. そこで前の記事の後ろに追跡してよい記事であるかを判断する役割を持つアルゴリズムとして後ろ向き K-Viterbi アルゴリズムを提案する. 後ろの内容の記事に対してのみ用いる. 語順の後ろから, モデルの後ろから尤度を推定していくが, 状態遷移確率, 初期状態確率, 出力記号確率を変更する必要はなく効果も K-Viterbi アルゴリズムと同様である.

(Step1’) 初期状態 S として可能な状態を P とする. この時のコスト $C_0(S) = [d_1(S, P_1), ..., d_k(S, P_i)]$,$(i = 1, ..., N)$

$m = 1, ..., M$ において,

(Step2’) 次の状態 Q と, その状態に入力する状態 P に対してコストを計算する.$C(P_i) = [k$ 個$] = [d_1(P_i, Q), ..., d_k(P_i, Q)]$

(Step3’) 各 Q において,$MAX_k[C_{m-1}(P_i) + d(P_i, Q)]$ を計算し, 大きい順に K 個の尤度をリスト $C_m(Q)$ に格納する.

(Step4’)Step2’〜Step3’ を繰り返す.

(Step5’) 最後に求めたコストに初期状態となる状態 Q の初期状態確率を掛け, 大きい順に K 個の尤度をリスト $C_m(Q)$ に格納する.

各記号は K-Viterbi アルゴリズムと同様である. 初期状態のコスト $C_0(S)$ の $d_k(S, P_i)$ は状態 P_i における出力記号確率である.

例えば図 3 のモデルにより, 記事の尤度上位 2 個の経路を得るとする. 記事の中身は” 山口, 周, 南市, 金峰, みたけ, し, 含む, 見つかっ,..., いう” となっている. 後ろ向き K-Viterbi アルゴリズムは語順の後ろから見ていくので, 最初の語は” いう” であり, モデルの’ 状態 3 と状態 2 と状態 1 からそれぞれ出力記号確率が 0.093...,0.0166...,0.01292... で出力されている. この時の各初期状態 S のコスト $C_0(S)$ は上位 2 個順に,0.093...,0.0166... である. さらに次の語へは状態 3 から状態 4, 状態 2 から状態 4 へ

の 2 つの遷移候補となる. これに対するコスト $C(P_i)$ は先程求めた状態 3 のコスト 0.093... を状態遷移確率と出力記号確率と共に掛け合わせる. 得られた値を大きい順に 2 個リストに格納すると, 最大で 0.0108... と 2 番目で 6.714E-4 となる. これを最後の語まで計算していき, 最後のコストの各値には対応する初期状態確率を掛ける.

以上の尤度計算を適用し,HMM によるトピックの局面追跡の過程を以下に設定する.

（StepA）2 種類のトピックの記事を各トピック内の時系列を保ちつつ直列にミックス. 特定対象のトピックの先頭記事を 1 件先頭に持ってくる.

（StepB）系列内の先頭の組から順に KL 情報量によりフィルタリングを行い, 閾値 ε_1 以下である類似した記事を選定.

$$\frac{KL(Dist(a_1)||Dist(a_1 \cup \ldots \cup \ a_j))+\ldots+KL(Dist(a_j)||Dist(a_1 \cup \ldots \cup \ a_j))}{j}$$

（StepC）系列内に残っている記事から 1 件抽出し,StepB で選定された記事集合と共にモデルを構築.

（StepD）構築したモデルで K-Viterbi アルゴリズム, および後ろ向き k-Viterbi アルゴリズムを用いて記事の尤度を推定する.

（StepE）StepB で選定された記事集合のパープレキシティ（PP）値の平均と StepC で抽出した記事の PP 値の差が閾値 ε_2 以下であれば類似しているとみなし選定.

（StepF）StepC に戻り系列内の次の記事を推定する.

j は比較している記事数を示す.StepB でフィルタリングを行うことにより, 語分布のみで抽出可能な類似した記事を選定する. 閾値 $\varepsilon_1, \varepsilon_2$ は幾つかの予備実験により決定する. 本研究では閾値 ε_1 を 0.012, ε_2 を 0.05 以下に設定する.

図 4 に例として, 図 3 のモデルと構築に用いた記事集合 7 件で記事 7 が他の 6 件と類似する記事であるかどうかを示す.

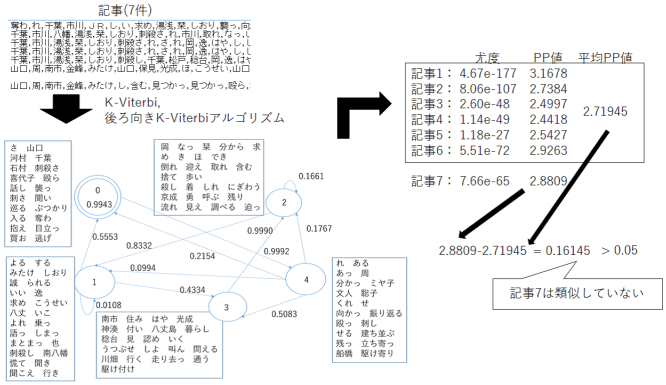

図 4　記事集合と構築されたマルコフモデルによる局面追跡

上述の局面追跡の過程により, 記事 7 の PP 値と平均 PP 値の差は閾値 0.05 以上であるため類似しない記事となる.

5. 実　　験

5.1 実験準備

関連研究である Yang 氏らの kNN による手法をベースラインとし, 提案手法である HMM を用いた尤度計算による手法を評価する. 各々によりトピックの追跡を行い, 追跡結果を視覚的に表現し適合率を比較する. 実験では「CD-毎日新聞 2013 データ集」の 1 月 1 日の先頭から同ジャンルであり, かつストーリーの流れがあり完結している「千葉・市川女性刺殺事件」の記事全 11 件と,「山口・周南放火殺人事件」の記事全 12 件をテストデータとして抽出する. ベースライン, 提案手法のどちらの場合においても扱う記事の中身は同じものとする.「CD-毎日新聞 2013 データ集」コーパスは予め MeCab により前処理をする. 各文章の単語に品詞付けをし, 固有名詞と動詞のみの自立語を抽出し, 不要語を除去する.

本研究では, 各トピック内の記事の時系列は保ちつつ直列にランダムに並べて開始する. また, 尤度の上位 K 個を常に候補として保有し, それぞれでパープレキシティ（PP）値を求める. 本実験では K=2 として行う.

各記事には各トピックのラベルが付与されているが, この正解ラベルを用いて各手法によるトピックの局面追跡の結果を評価する.

ベースライン及び提案手法において予備実験により求めた閾値を用いる. ベースラインでは学習データを各トピックの記事集合の先頭から 1 件ずつ割り当て, 閾値は 0.2,K=3 と設定する.

評価は, 正規化した Detection Cost function(注1) を用いる [6]. 対象のストーリー「事件発生」,「逮捕」,「供述-1」,「被害者の人柄」,「供述-2」,「起訴」の各事象での記事を何件追跡出来なかったかを示す未検出率と, 誤って追跡した他トピックの記事件数は何件であるかを示す誤検出率に基づいて算出する. 値が 0 に近いほどシステムの性能が高いことを示している. さらに適合率と再現率を求め評価する.

5.2 実験結果

追跡対象とするトピックを「千葉・市川女性刺殺事件」とする. ベースライン及び提案手法によるトピックの局面追跡の未検出率と誤検出率, 適合率と再現率を表 1 に示す. 図 5,6 には各手法により追跡した記事と正解ラベル, また記事の中身の一部を示す.

表 1　各手法における性能評価値

	ベースライン	提案手法
正解/結果	件数	件数
発生 (1 件)	1	1
逮捕 (3 件)	1	1
供述-1(1 件)	0	1
被害者の人柄 (1 件)	0	0
供述-2(4 件)	3	2
起訴 (1 件)	0	1
抽出した他トピックの記事 (12 件)	4	2
未検出率 (%)	54.545... (6/11)	45.454... (5/11)
誤検出率 (%)	33.333... (4/12)	16.666... (2/12)
Norm Detection Cost	6.6107...	4.3332...
適合率（%）	55.555...(5/9)	75(6/8)
再現率（%）	45.454...(5/11)	54.545...(6/11)

(注1)：未検出率と誤検出率に基づいてシステムの性能を評価する関数. 式は $C_{Det} = C_{miss} \times P_{miss} \times P_{target} + C_{false} \times P_{false} \times (1 - P_{target})$ で与えられる. 正規化の式は,$(C_{Det})_{Norm} = C_{Det}/Min(C_{miss} \times P_{target}, C_{false} \times (1 - P_{target}))$ である.C_{miss},C_{false} は論文中の表記になぞらえてそれぞれ 10, 1 で与えている.P_{miss} は未検出率,P_{false} は誤検出率,P_{target} はコーパス中に含まれる正解トピックの記事集合の割合である.

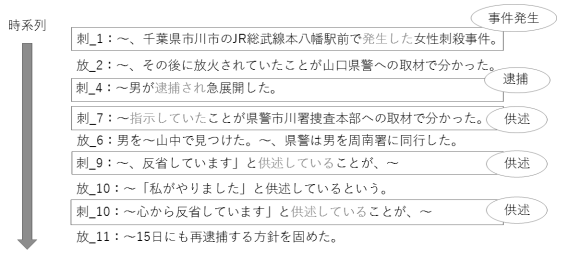

図 5　ベースラインによるトピックの追跡結果

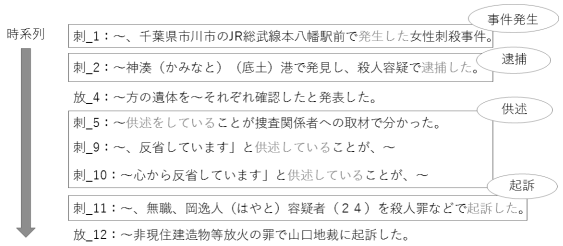

図 6　提案手法によるトピックの追跡結果

表 1 で示すように, ベースラインの場合では未検出率が 54.5 % (6/11) であり, 誤検出率は 33.3 % (4/12) である.Detection Cost は 6.6107... である. 抽出した記事は 9 件であり, 内 5 件は「千葉・市川女性刺殺事件」の記事であり, 残り 4 件は「山口・周南放火殺人事件」に関する「容疑者起訴」の記事である. 局面追跡の結果は,「事件発生 (千葉・市川の記事 1 件) →事件発生（山口・周南の記事 1 件）→犯人逮捕 (千葉・市川の記事 1 件) →供述 (千葉・市川の記事 1 件) →犯人逮捕（山口・周南の記事 1 件）→供述（千葉・市川の記事 1 件）→供述（山口・周南の記事 1 件）→供述（千葉・市川の記事 1 件）→容疑者再逮捕 (山口周南の記事 1 件)」となる. 提案手法の場合の未検出率は 45.5 % (5/11) であり, 誤検出率は 16.7 % (2/12) である.Detection Cost は 4.3332... である. 抽出した記事は 8 件であり, 内 6 件は「千葉・市川女性刺殺事件」に関する記事, 残り 2 件は「山口・周南放火殺人事件」に関する「容疑者起訴」と「捜査」の記事である. 局面追跡の結果は,「事件発生 (千葉・市川の記事 1 件) →犯人逮捕 (千葉・市川の記事 1 件) →捜査 (山口・周南の記事 1 件) →供述 (千葉・市川の記事 3 件) →容疑者起訴 (千葉・市川の記事 1 件) →容疑者起訴 (山口・周南の記事 1 件)」となる.

提案手法により追跡した記事でマルコフモデルを構築する. 図 7 に構築したマルコフモデルを示す. 出力語は他の状態から出力していない語のみを示している. 状態名は, 各状態に付随する出力語が集中して多く出現する記事の内容に基づき決定している.

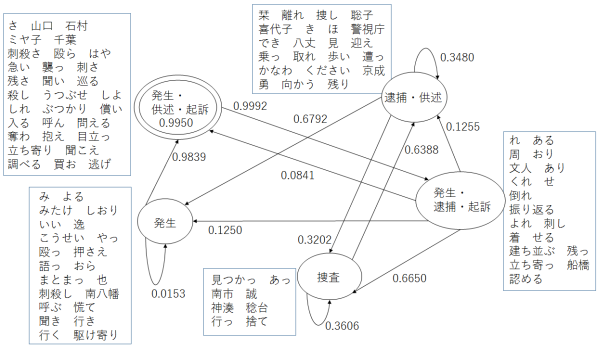

図 7　提案手法により追跡した記事のマルコフモデル

提案手法において, ベースラインで追跡できていない「起訴」に関する内容の記事を追跡しており, 追跡対象であるトピック「千葉・市川女性刺殺事件」のストーリーの変遷を違和感無く捉えている.

5.3　考　　察

5.3.1　各手法の再現率

提案手法では 6 件の正解トピックの記事を抽出し, 再現率は 54.54.. %である. また, 提案手法により追跡結果であるトピックの内容の変遷がベースラインより改善している. また, ベースラインで追跡しきれていない「起訴」に関する記事を追跡出来ている. この原因として, トピックの構造として構築したマルコフモデルにあると考える. 図 8 に追跡対象のトピックのマルコフモデルを示す.

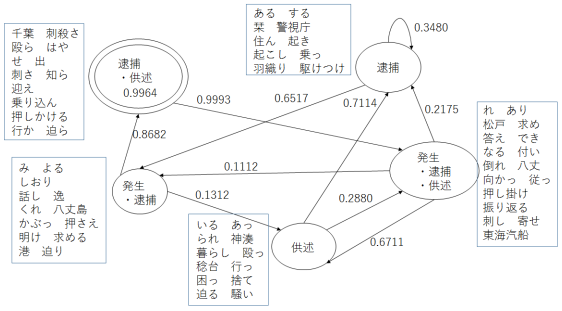

図 8　正解のトピックのマルコフモデル

提案手法により構築したマルコフモデルは正解のモデルと比較するとほぼ状態名が同じであり, 正解トピックの内容の変遷を構築している. また, 各状態の出力記号がほぼ全て正解トピックの記事の語に属している. このモデルにより推定された各記事の PP 値をまとめた表 2 を以下に示す. 各ラベルは各トピック名と番号が振られている. 差が閾値以下の記事には下線が引かれている.

これらの原因により, 正解トピックの記事が選定され再現率が向上し, 内容の変遷が改善したと考えられる.

5.3.2　各手法における誤検出率

誤検出率は提案手法では 16.7 % (2/12) であり, 提案手法のフィルタリングの過程で 2 件抽出している. この原因として, 予備実験により求めた閾値によるものと考えられる. 表 3 に一部の各閾値と, 抽出される記事と類似度を示す.

表 2　提案手法による各記事の PP 値

k=2, 閾値：0.05											
1 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南			1-周南	平均	差
1-PP 値	2.078527985	1.734456688	1.451737874	1.480746562	1.561471534	1.746526851			1	1.675577916	0.675577916
2-PP 値	2.07960942	1.735133921	1.454990313	1.48382782	1.594916409	1.748490572			1	1.682828076	0.682828076
2 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南			2-周南	平均	差
1-PP 値	3.1678969	2.73844672	2.499749457	2.441880777	2.542764137	2.926392123			2.881038878	2.719521686	0.161517192
2-PP 値	3.188159355	2.746918137	2.500719727	2.475342899	2.591735655	2.958070899			2.927718479	2.743491112	0.184227367
3 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南			3-周南	平均	差
1-PP 値	3.911065496	3.001982644	2.542555537	2.532693245	2.64521437	3.335302439			3.18797105	2.994802288	0.193168762
2-PP 値	3.915252168	3.014604627	2.56959152	2.563667345	2.707017881	3.056993687			3.194263555	2.971187871	0.223075684
4 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南			3-市川	平均	差
1-PP 値	3.319597541	2.869100708	2.549796656	2.54740159	2.562376989	3.138191746			3.099702152	2.831077538	0.268624613
2-PP 値	3.32712179	2.869370694	2.553978382	2.595353343	2.686681913	3.163489744			3.128674972	2.865999311	0.262675661
5 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南			4-市川	平均	差
1-PP 値	3.328419066	2.897115068	2.546987597	2.530027903	2.547723285	3.163541404			3.114855714	2.83563572	0.279219994
2-PP 値	3.333531817	2.900332244	2.565742674	2.547674148	2.586598556	3.175401785			3.13132109	2.851546871	0.27977422
6 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南			4-周南	平均	差
1-PP 値	3.450070263	2.836898909	2.50538	2.498866076	2.543737984	3.029135723			2.862720962	2.810681492	0.05203947
2-PP 値	3.453774952	2.855988822	2.528789807	2.520980706	2.65423756	3.045084659			2.873076135	2.843142751	0.029933384
7 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南	4-周南		5-市川	平均	差
1-PP 値	3.545251877	3.057047524	2.589711267	2.629067895	2.485754622	3.140404927	2.968552259		2.94895619	2.916541481	0.032414708
2-PP 値	3.54730834	3.058228065	2.59173959	2.631018626	2.495080913	3.147910354	2.994458192		2.949813184	2.923677726	0.026135458
8 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南	4-周南	5-市川	5-周南	平均	差
1-PP 値	3.59499494	3.06453146	2.587402664	2.63401163	2.553265416	3.22081403	2.873628988	3.027194967	3.160321028	2.944480512	0.215840516
2-PP 値	3.596027564	3.06580972	2.589591597	2.636122664	2.557154826	3.224782383	2.878006946	3.027926226	3.16307107	2.946927741	0.216143329
9 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南	4-周南	5-市川	6-市川	平均	差
1-PP 値	3.604316139	3.06940139	2.590503148	2.632189867	2.541230447	3.205083102	2.880984103	3.014868116	5.003878613	2.942322039	2.061556574
2-PP 値	3.604316139	3.06940139	2.599525541	2.64087415	2.557184046	3.219905953	2.887884055	3.016447011	5.008530167	2.949442286	2.059087882
10 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南	4-周南	5-市川	7-市川	平均	差
1-PP 値	4.117761226	3.233030558	2.726108651	2.719426509	2.607672544	3.560545635	3.183374817	3.133962331	3.559633675	3.160235284	0.399398391
2-PP 値	4.120837543	3.233978271	2.732068666	2.722644069	2.695044297	3.574277253	3.189043192	3.137260478	3.563474569	3.175644221	0.387830348
11 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南	4-周南	5-市川	6-周南	平均	差
1-PP 値	3.808852288	3.167533593	2.661204354	2.64317868	2.633367032	3.102562701	2.956608579	3.148319732	3.323764105	3.01520337	0.308560735
2-PP 値	3.811141453	3.168964686	2.669049557	2.656382412	2.637712242	3.123425183	2.96485655	3.149390147	3.324195449	3.022615279	0.301580171
12 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南	4-周南	5-市川	8-市川	平均	差
1-PP 値	4.063934391	3.214851077	2.625451219	2.641054811	2.725990365	3.594594793	3.247234557	3.107155568	3.21573573	3.152533348	0.063202383
2-PP 値	4.06476014	3.217576088	2.627558093	2.654669955	2.810341827	3.5997716	3.254062244	3.109137441	3.219285711	3.167234674	0.052051038
13 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南	4-周南	5-市川	7-周南	平均	差
1-PP 値	3.791662396	3.020386772	2.687049963	2.650945341	2.661408771	3.299299277	3.143811218	3.175171929	3.597046645	3.053716958	0.543329687
2-PP 値	3.792634963	3.024414589	2.695467402	2.681530314	2.726862707	3.300319433	3.145343594	3.177694251	3.597433171	3.068033407	0.529399765
14 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南	4-周南	5-市川	8-周南	平均	差
1-PP 値	4.306145434	3.304064385	2.595207982	2.689184219	2.571440481	3.444210405	3.085270947	3.240042744	3.662901489	3.154445825	0.508455664
2-PP 値	4.308183557	2.277557866	2.609389509	2.697987139	2.672228437	3.456156767	3.123300146	3.241177667	3.670293928	3.048247636	0.622046292
15 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南	4-周南	5-市川	9-周南	平均	差
1-PP 値	3.641310911	3.070952446	2.744777432	2.662032196	2.667102253	3.172548212	2.934641315	3.176271965	3.293369682	3.008704591	0.284665091
2-PP 値	3.647715938	2.177057995	2.781377421	2.709568291	2.7082955	3.202769174	2.954058242	3.17667293	3.298411204	2.919689436	0.378721768
16 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南	4-周南	5-市川	10-周南	平均	差
1-PP 値	4.15354914	3.211257649	2.678730712	2.693741422	2.815669687	3.192528355	3.207793039	3.2006679	4.083552119	3.144242238	0.939309881
2-PP 値	4.160047788	2.234797994	2.750208316	2.698930219	2.872927645	3.198198772	3.212763361	3.206026622	4.086266696	3.04173759	1.044529106
17 回目	1-市川	2-市川	9-市川	10-市川	11-市川	12-周南	4-周南	5-市川	11-周南	平均	差
1-PP 値	4.17599415	3.19078039	2.554657864	2.591940053	2.8332368	3.010507585	3.082045878	3.19194106	3.303722023	3.078887973	0.224834051
2-PP 値	4.177876348	2.225310428	2.619565769	2.609155534	2.873178409	3.020857092	3.082890878	3.192256801	3.319122978	2.975136407	0.34398657

表 3　各閾値による抽出された記事とその類似度

閾値 0.01			閾値 0.012		
	比較する記事	類似度 (KL 情報量)		比較する記事	類似度 (KL 情報量)
1 回目	刺-1, 放-1	0.021085951	1 回目	刺-1, 放-1	0.021085951
2 回目	刺-1, 放-2	0.013214638	2 回目	刺-1, 放-2	0.013214638
3 回目	刺-1, 刺-2	0.008882086	3 回目	刺-1, 刺-2	0.008882086
4 回目	刺-1, 刺-2, 放-3	0.020165253	4 回目	刺-1, 刺-2, 放-3	0.020165253
5 回目	刺-1, 刺-2, 刺-3	0.012190679	5 回目	刺-1, 刺-2, 刺-3	0.012190679
6 回目	刺-1, 刺-2, 放-4	0.013786736	6 回目	刺-1, 刺-2, 放-4	0.013786736
7 回目	刺-1, 刺-2, 刺-5	0.012829681	7 回目	刺-1, 刺-2, 刺-5	0.012829681
8 回目	刺-1, 刺-2, 放-5	0.014469782	8 回目	刺-1, 刺-2, 放-5	0.014469782
9 回目	刺-1, 刺-2, 刺-6	0.013364345	9 回目	刺-1, 刺-2, 刺-6	0.013364345
10 回目	刺-1, 刺-2, 刺-7	0.01461821	10 回目	刺-1, 刺-2, 刺-7	0.01461821
11 回目	刺-1, 刺-2, 放-6	0.017462573	11 回目	刺-1, 刺-2, 放-6	0.017462573
12 回目	刺-1, 刺-2, 刺-8	0.012732278	12 回目	刺-1, 刺-2, 刺-8	0.012732278
13 回目	刺-1, 刺-2, 放-7	0.017364093	13 回目	刺-1, 刺-2, 放-7	0.017364093
14 回目	刺-1, 刺-2, 刺-9	0.010989098	14 回目	刺-1, 刺-2, 刺-9	0.010989098
15 回目	刺-1, 刺-2, 放-8	0.015315333	15 回目	刺-1, 刺-2, 刺-9, 放-8	0.014939441
16 回目	刺-1, 刺-2, 放-9	0.019816968	16 回目	刺-1, 刺-2, 刺-9, 放-9	0.018760831
17 回目	刺-1, 刺-2, 放-10	0.016737215	17 回目	刺-1, 刺-2, 刺-9, 放-10	0.01625942
18 回目	刺-1, 刺-2, 刺-10	0.010969145	18 回目	刺-1, 刺-2, 刺-9, 刺-10	0.010330202
19 回目	刺-1, 刺-2, 放-11	0.016511058	19 回目	刺-1, 刺-2, 刺-9, 刺-10, 放-11	0.014270153
20 回目	刺-1, 刺-2, 刺-11	0.011012479	20 回目	刺-1, 刺-2, 刺-9, 刺-10, 刺-11	0.010174216
21 回目	刺-1, 刺-2, 放-12	0.014130583	21 回目	刺-1, 刺-2, 刺-9, 刺-10, 刺-11, 放-12	0.011855117
抽出された記事	刺-1, 刺-2			刺-1, 刺-2, 刺-9, 刺-10, 刺-11, 放-12	

表 3 から, 閾値の変化により類似するとして抽出される記事が異なる. 閾値 0.01 の場合では正解トピックの記事のみ抽出されているが, 閾値 0.012 では他トピックの記事が閾値を下回り抽出されている. これらの原因により, 誤って他トピックの記事が抽出されたと考えられる.

6. 結　　論

本研究では, 新たな尤度計算の方法である K-Viterbi アルゴリズム及び後ろ向き K-Viterbi アルゴリズムを提案し, 推定した尤度によるトピックの局面追跡の手法を提案した. 隠れマルコフモデルを適用することで学習データを必要とせず, 自動的に特定のトピックに対する内容の変遷を捉えることを提案した. 結果としてベースラインより適合率が 20 %, 再現率が 9 %向上し, 未検出率が 9 %, 誤検出率は 17 %改善された. また, マルコフモデルを構築することでトピックのストーリーの構造を図式化することで, 状態の遷移過程が正解トピックのストーリーの構成と見て取れることを確認した. ベースラインに比べて違和感のない内容の変遷を捉え, かつ容疑者の起訴などの重要な記事を抽出することが可能となった.

謝　　辞

本研究の一部は Wai Lam 教授 (Chinese University of Hong Kong) の助言を受けて行われた. ここに感謝の意を表します.

文　　献

[1] James Allan, Ron Papka, and Victor Lavrenko, “On-line New Event Detection and Tracking”, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 1–9, 1998
[2] Andrew D. Brown, Geoffrey E. Hinton, “Products of Hidden Markov Models”, Gatsby Computational Neuroscience Unit, pp. 1–8, 2001
[3] Daniel Jurafsky, James H. Martin, “Hidden Markov Model”, Speech and Language Processing, 2016
[4] James Allan, Jaime Carbonell, Geroge Doddington, Jonathan Yamron, and Yiming Yang, “Topic Detection and Tracking Pilot Study Final Report”, In Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, pp. 194–218, 1998
[5] Jeff A. Bilmes, “A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gausian Mixture and Hidden Markov Models”, ICSI, pp. 1–15, 1998
[6] Jonathan G. Fiscus, George R. Doddington, “Topic Detection and Tracking Evaluation Overview”, Topic detection and tracking, pp. 17–31, 2002
[7] Marek Gajewski, Janusz Kacprzyk, Slawomir Zadrozny, “TOPIC DETECTION AND TRACKING: A FOCUSED SURVEY AND A NEW VARIANT”, ZNWI, pp. 1–16, 2014
[8] Mrs. Lavanya. S, Kavipriya. R, “A Survey on Event Detection in News Streams”, IJCST, pp. 1–3, 2014
[9] Yiming Yang, Jaime G. Carbonell, Ralf D. Brown, Tom Pierce, Brian T. Archibald, Xin Lin, “Learning approaches for Detecting and Tracking News Events”, IEEE Journal, pp. 32–43, 1999
[10] 井庭 崇, “探索的モデリング”, pp. 5–11, 2004
[11] 北田 剛士, 風間 一洋, 榊 剛史, 鳥海 不二夫, 栗原 聡, 篠田 孝祐, 野田 五十樹, 斎藤 和己, “Twitter のトピック変遷の可視化法の提案”, DEIM, pp. 1–8, 2015
[12] 若林 啓, 三浦 孝夫, “HMM を用いた文書における事象系列の推定”, DBSJ, pp. 1–4, 2007