確率過程モデルを用いた時系列データからの知識獲得に関する研究

若林, 啓 / WAKABAYASHI, Kei

(発行年 / Year)
2009-03-24
(学位授与年月日 / Date of Granted)
2009-03-24
(学位名 / Degree Name)
修士(工学)
(学位授与機関 / Degree Grantor)
法政大学 (Hosei University)

2008年度　修 士 論 文

確率過程モデルを用いた
時系列データからの知識獲得に関する研究

STUDIES ON KNOWLEDGE DISCOVERY FROM TIME-SERIES DATA USING STOCHASTIC MODELS

指導教官　三浦 孝夫 教授

法政大学大学院工学研究科
電気工学専攻修士課程

07R3138　若林 啓
Kei WAKABAYASHI

目 次

第1章　序論

1.1　問題の背景

近年, 計算機環境の充実に伴い, 大量の情報を高速に扱うことが可能となり, 様々なデータが電子化され計算機上で管理されている. また, インターネット技術の発達により, 情報を電子化することによる利便性の向上や人間の作業量の軽減がますます進んでおり, 様々な種類の情報が電子化されるようになった. しかし, 膨大なデータが利用可能になった一方で, データが表現する内容をユーザーが人手で解釈し, 知識を獲得することはますます困難になってきている. データマイニングは, 大規模なデータベースから自動的に有用な知識を発見する技術である. これまでに提案されている多くのデータマイニング技術は, 静的な関係データベースを対象にして, 頻出するパターンやデータの相関関係を発見する問題を扱っている.

一方, 近年では, 情報の処理や伝達の高速化に伴い, リアルタイムに大量の情報を更新するデータベースが実現している. このようなデータベースにおいて, データの時系列的特徴を含めたパターン発見は有用な技術である. データの時系列的特徴とは, データの発行された時刻の前後関係に関連したパターンを指す. 例えば, ある種類のデータが発行されたとき, 次の時刻に発行されるデータがある特定の種類に限られるような規則を発見することは, 事象の解析やデータの予測で有用である.

静的な関係データベースを対象としたデータマイニング手法は, 個々のデータの関係が集合によって定義されることを前提としており, 時系列的特徴に基づくパターン発見にはそのまま適用できない. これは, データの集合を対象としたパターン発見とは異なり, 走査する必要のあるデータの関係の数が時刻間の関係を考慮することで爆発的に増加するためである. 例えば, ある時刻のデータが直前の時刻のデータにのみ依存していると仮定しても, 一般に各時刻のデータは複数の属性のデータを含むベクトルであるため, 走査する必要のあるデータの関係は属性の組み合わせの数だけ存在し, 膨大である.

このことから, 時系列的特徴に基づくパターン発見問題では, 実現可能な計算量で効果的なパターン発見を行うアルゴリズムの実現が求められている. 現在, 時系列的特徴に基づくパターン発見問題は様々な分野で研究が行われており, 時系列データ予測や異常値検出, 時系列的特徴に基づく系列データ分類などの応用が期待されている. また, 画像認識の分野では手書き文字認識や監視カメラにおける異常行動検出, 音声認識の分野では音声の時系列的特徴に基づく話者特定といった問題に関連して時系列データのパターン発見の研究が進められている. 自然言語処理の分野においては, ウェブ上の文書や新聞記事といったデータも時系列的特徴を持つため, 文書内容の予測や文書系列の

分類といった応用も考えられる.

1.2 扱う問題

本研究では, 時系列的特徴に基づくパターン発見問題に関連して, 時系列データのモデル化の手法を提案する.

従来の静的な関係データベースにおけるパターン発見手法は, データ間の直接の関係を走査することで特徴的なパターンの抽出を行う. しかし, この手法は特徴的な属性集合を発見するために, 膨大な属性の組み合わせを考慮する必要がある.

本研究では, 隠れ状態を含む確率過程モデルを用いて時系列データをモデル化を行い, 対象の時系列データの特徴をモデルのパラメタとして抽出する手法を提案する. 本手法ではデータ間の関係を直接考えず, 非観測である隠れ状態の確率的な遷移に基づいてデータ間の関係を与える. これにより, 膨大な組み合わせの走査を行うことなく時系列的特徴を抽出できることを以下の問題を通して論じる.

1.2.1 時系列的特徴に基づく新聞記事系列の分類

時系列データとして文書の系列が与えられたとき, その時系列的特徴を抽出することは, 文書内容の把握や分類, 予測に有用である. しかし, 文書データは同じ内容でも表現の違いがあるなど, その時系列パターンを形式化することは容易ではない. ここでは, 文書データとして新聞記事を用いることにより, 一連の事件を時系列的特徴に基づいて分類する手法について考える.

本研究では, 日本語の特徴により, 新聞記事から事件の進展を最もよく表現していると考えられる各文章末の動詞を抽出し, 事件を表現する系列データとみなす. しかし, 抽出した動詞の直接の遷移関係をパターンとすると, 効果的な分類が行えない. これは例えば, 犯人が捕まったことを述べている記事で「逮捕した」と「拘束した」などのように複数の異なる表記の動詞が考えられる. ここでは動詞の遷移ではなく, 隠れた「事象」の遷移に基づいて事件の時系列的なパターンを形式化する.

本研究では, 隠れ状態を含む確率過程モデルを用いて事件のモデル化を行い, 事象の系列の特徴をモデルのパラメタとして抽出する. 複数の種類の事件に対し, それぞれ独立に確率過程モデルの学習を行い, 与えられた新聞記事系列について最も尤度を大きくするモデルの事件に分類する手法を提案する.

1.2.2 新聞記事からの詳細な時系列的特徴の抽出

新聞記事の要約として各文章末の動詞を利用することは, 時系列的特徴の抽出に有用である. しかし, 動詞のみを記事の特徴とすると, 非常に単純な事象系列パターンしかモデルに反映させることができないという問題がある. 例えば, 新事実の発覚に関する

事象は, 動詞だけを用いると「分かった」という情報しか得られないが, 発覚した内容は事件の進展の判別に効果的であると考えられる.

ここでは, 動詞の共起語を文書の特徴として用いることで, 詳細な時系列的特徴をモデルに学習させることを目指す. 動詞の遷移パターンと, それぞれの動詞に伴って出現した共起語の特徴に基づいて確率過程モデルの学習を行い, 事件の分類を行う手法を示す.

1.2.3 事象系列に基づく時系列データ予測

時系列データ予測は, 過去の履歴データに基づいて未来のデータの予測を行う技術であり, 商店における効果的な在庫管理や企業の客観的な意思決定に有用である. ここでは各時刻で, 数値の多次元ベクトルデータを観測値として与える.

多くの時系列データ予測の手法は, 前の時刻のデータの値を, 線形回帰などの方法により直接予測値の推定に用いる. しかし, この手法は事象に基づいて発生する観測値の予測を効果的に行うことができない. 事象に基づく時系列データとは, 風速の観測データが「台風の接近」という事象に基づいて発生するように, 直接観測値に現れない隠れた事実に依存した時系列データを意味する. このような時系列データでは, 前の観測値に直接依存しない構造を持った予測を行う必要がある.

また, 特にデータストリーム環境では, 途中でデータの分布が動的に変化することが想定される. このことから, 学習データを用いてあらかじめ時系列パターンを学習したとしても, 予測の最中に未知の事象が発生することを考えなければならない.

本研究では, 隠れ状態を事象に対応させた隠れマルコフモデルを用いて時系列的特徴のモデル化を行い, 隠れ状態の遷移に基づいて観測値の予測を行う. また, 差分型の学習アルゴリズムを適用し, 予測の最中に新たな観測データの傾向をモデルに反映させることにより, データストリームにおける動的なデータの分布の変化に適応的な予測を行うことを目指す.

1.3 論文の構成

本研究では，以上の問題について以下の構成で論じる．第 2 章では，HMM を用いた文書の事象系列に基づいたトピック分類手法について論じる．第 3 章では，共起語を利用した文書のトピック推定について論じる．第 4 章では，差分型 HMM を用いたデータストリームにおける時系列データ予測手法を提案する．第 5 章で結論とする．

1.4 発表論文

1. 若林啓, 三浦孝夫: “HMM を用いた文書における状況系列の推定”，データ工学ワークショップ (*DEWS*), 2007.

新聞記事の系列を一連の事件とみなして, 事象系列に基づくトピック推定を行う. 隠れマルコフモデルを用いた新しい文書モデルを提案し, 実験によりその有用性を検証する.

2. 若林啓, 三浦孝夫: "Identifying Event Sequences using Hidden Markov Model", *12th International Conference on Applications of Natural Language to Information Systems* (*NLDB*), pp.84-95, 2007.
新聞記事の系列を一連の事件とみなして, 事象系列に基づくトピック推定を行う. 隠れマルコフモデルを用いた新しい文書モデルを提案し, 実験によりその有用性を検証する.

3. 若林啓, 三浦孝夫: "HMM を用いた文書における状況系列の推定", 日本データベース学会 Letters (*DBSJ Letters*) Vol.6, No.3, pp.17-20, 2007.
新聞記事の系列を一連の事件とみなして, 事象系列に基づくトピック推定を行う. 隠れマルコフモデルを用いた新しい文書モデルを提案し, 実験によりその有用性を検証する.

4. 若林啓, 三浦孝夫: "共起語を利用した事象系列に基づくトピック推定", データ工学ワークショップ (*DEWS*), 2008.
文書の特徴語として, 日本語の特性に基づいた重要語とその共起語を用いて事件のトピック推定を行う. 確率過程モデルによる事件のモデル化を行い, 実験により有用性を検証する.

5. 若林啓, 三浦孝夫: "Topics Identification Based on Event Sequence Using Co-occurrence Words", *13th Intn'l Conf. on Applications of Natural Language to Information Systems* (*NLDB*), pp.219-225, 2008.
文書の特徴語として, 日本語の特性に基づいた重要語とその共起語を用いて事件のトピック推定を行う. 確率過程モデルによる事件のモデル化を行い, 実験により有用性を検証する.

6. 若林啓, 三浦孝夫: "共起語を利用した事象系列に基づくトピック推定", 日本データベース学会論文誌 (*Journal of the DBSJ*) Vol.7, No.1, pp.79-84, 2008.
文書の特徴語として, 日本語の特性に基づいた重要語とその共起語を用いて事件のトピック推定を行う. 確率過程モデルによる事件のモデル化を行い, 実験により有用性を検証する.

7. 若林啓, 三浦孝夫: “差分型 HMM を用いたデータストリームにおける時系列データ予測”, データ工学と情報マネジメントに関するフォーラム (*DEIM*), 2009.
隠れマルコフモデルによるパターンの学習に基づいて, 事象系列に基づく時系列データ予測を行う. また, 差分型学習を適用することで動的なデータの分布の変化に適応可能な予測手法を示し, 実験によりその有用性を示す.

第2章　HMMを用いた文書における状況系列の推定

本稿では, 文書で表現されたトピックを分類する手法を提案する. これまでに文書をモデル化する手法については多く論じられてきたが, 直接文書の内容を扱った研究は少ない. 一連のトピックは状況の系列によって表現できる. 本研究では, 一連の新聞記事からの状況の系列の推定を HMM によるタグ付け問題として扱い, 実験により手法の妥当性を示す.

2.1　前書き

近年, 計算機上で利用可能な文書データの増加に伴い, より高度な知識処理技術が必要とされている. この現状を背景にして, 文書分類技術に関する研究が盛んに行われている. 文書分類技術は, 一般的に文書データを出現単語ベクトルにモデル化することで分類を行う. しかしこのベクトルモデルではその文書が述べているトピックを扱うことは難しい. 本研究では文書そのものではなく, その文書が表現しているトピックを対象にモデル化を考える. 本稿の目的は, 文書分類ではなくトピックの分類である.

トピックを扱う代表的なアプローチの一つに, *Topic Detection and Tracking*（TDT）がある [1]. TDT では, トピックは事象（event）によって特徴付けられる. 事象とは, 位置的, 時間的に特定の, 個々の発生した事実を意味する. TDT の Event tracking タスクでは, ある事象に関して述べている文書を逐次的に分類する [5].

本研究ではトピックを事象の系列と考えることで隠れマルコフモデル (Hidden Markov Model, HMM) を適用し, トピックを形式化することを考える. 事象系列を考慮した分類手法は過去にあまり積極的な提案はなされていない. というのは，決定木や SVM といった従来の分類手法では, ベクトル分類に帰着させることが多く，系列情報を反映させることは容易ではない.

本研究では事象系列を分類するための新しい手法として, 確率過程に基づいた文書の表現モデルを提案する. 確率過程は事象の系列をモデル化したものであり, 事象間の遷移を扱うことができる. このため, 文書を確率過程と考えることで, トピック分類を行うことができる.

文書を確率過程とみなす研究には,Barzilay らがある [2]. ここでは特定の種類の文書を対象として, 文書構造を HMM を用いて推定している. HMM の出力シンボルとして,

文章の bigram 確率を導入している点で興味深い．しかし, 本研究では文書そのものの構造ではなく, 文章が表現しているトピックを対象にしているため, この手法をそのまま適当することはできない.

文書の内容を確率過程で直接扱う研究には, 柴田らがある [4]．ここでは料理番組のナレーションを対象にして, 文書中の任意の位置でどのような調理を行っている状況であるかを HMM を用いて推定している．ただし, この研究は単一のトピックに限定した状況の推定を目的としているため, 複数の種類のトピックは扱わない.

2 章では本研究が扱うトピックの分類について述べる．3 章では隠れマルコフモデルについて説明する．4 章で具体的な事象推定のアルゴリズムの説明を行う．5 章で実験結果を示し,6 章で結びとする.

2.2 事象系列によるトピックの分類

ここでは, 本稿で提案する手法のアイデアを例を用いて述べる．本論文中では,「状況」(「事象」とも言う）は, 位置的, 時間的に特定される個々の事柄, 発生した事実を意味する．「出来事」(「トピック」とも言う）は, 一連の事件やテーマに関する状況の系列を意味する．単に状況の系列（事象の系列）と言う場合, 特に一連の事件やテーマとは無関係な状況の系列を意味する.

２つのトピックが似ているかどうかは,「似ている」の解釈によって判断が分かれる．このためトピックの分類方法は一般に一意ではない．本研究ではトピックを事象の系列と考えるため, 事象の系列が似ているトピックを「似ている」ものとする．例えば, 東京で起きた強盗事件と広島で起きた強盗事件は, 場所も犯人も盗品も違う．しかし, どちらも盗まれた, 指名手配された, 犯人が逮捕された, と事象の系列が似ているならば, 両トピックは似ているとする.

図 2.1 は, 本研究で考えるトピックの推定の例である．いま, 図中の左側に示されているような,「トピックを表現している文書」が与えられている．本稿では, 新聞記事の第一段落を日付順に連結し時系列に並べた文書を考える．この例では, ある殺人事件に関する新聞記事を連結させたものである．この文書は「殺人事件」というトピックを表現している文書となっている.

図中の右側に示しているのは, このトピックでたどっている「事象の系列」である．文書の解読から, ある人物の死亡の発見という事象から始まる．次に, 警察の調べによって不審な人物の手がかりが明らかになる事象が続く. 最後に, 容疑者が逮捕されるという事象が得られる．これらがこのトピックの構成である．無論このトピックは,「殺人事件」という特性に依存する．ある人物の自殺に関するトピックの場合, 推定される事象として，最初に自殺の発覚, 次は自殺の原因が明らかになる等の系列になるであろう.

殺人事件トピックの最後では逮捕という事象が見られるが, 自殺トピックとの関連からは, 逮捕された容疑者が獄中で自殺を図ることが考えられる．しかし, ここではすでに自殺発覚の事象があるため, 事象の順序としては不自然である．即ち, 殺人事件トピックは自殺トピックとは通常両立しない系列を有している.

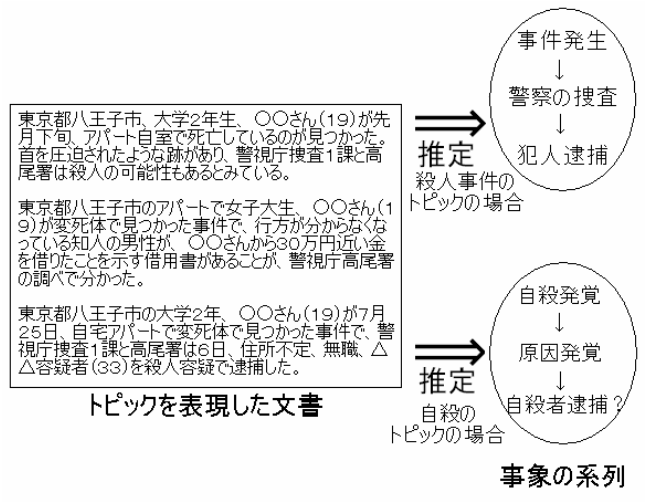


図 2.1: 文書からのトピックの推定

事象系列が, トピックの種類ごとに存在すると考えられることから，本研究では,（事象系列に基づく）トピックを推定することによって当該トピックの分類を行うことができることを論じる.

2.3 隠れマルコフモデル

隠れマルコフモデルは, 確率的に遷移する内部状態をもつオートマトンである. 内部状態は単純マルコフ過程に従って遷移する. 通常, 内部状態は直接観測できない. その代わりそれぞれの内部状態は, 一つの観測可能なシンボルを確率的に出力する.

2.3.1 モデルの定義

隠れマルコフモデルは次の 5 つのパラメータによって定義される [3].

(1) $Q = \{q_1, \cdots, q_N\}$：状態の有限集合

(2) $\Sigma = \{o_1, \cdots, o_M\}$：出力シンボルの有限集合

(3) $A = \{a_{ij}\}$：状態遷移確率分布
a_{ij} は状態 q_i から状態 q_j への遷移確率である.

(4) $B = \{b_i(o_t)\}$：シンボル出力確率分布
$b_i(o_t)$ は状態 q_i でシンボル o_t を出力する確率である.

(5) $\pi = \{\pi_i\}$：初期状態確率分布
π_i は状態 q_i が初期状態である確率である.

本研究では, 状態は事件発生, 容疑者の逮捕, 自殺発覚などの事象の種類に対応する. 状態集合 Q はトピックの種類によって異なる集合をもつ. 出力シンボルは観測可能な情報であり, 文書に該当するが, 次章で述べる特徴的な単語の抽出によって得る単語のみをシンボルとする.

状態遷移確率分布 A はある状態から次の状態へ遷移する確率であるため, ある事象が起きた後, 次に起こる事象の確率分布である. シンボル出力確率分布 B は, ある事象が起きたとき, 文書中にどのような単語が出現するかを表す確率分布である. 初期状態確率分布 π は, 最初に起きる事象の確率分布である.

2.3.2 状態列の推定

隠れマルコフモデルは, 観測したシンボル列から, 隠れた内部状態列を推定する目的で用いることが多い. モデルのパラメータに基づいて, 与えられたシンボル列に対して最適な内部状態列を求める問題を, 隠れマルコフモデルの復号化問題と呼ぶ. Viterbi アルゴリズムは, 復号化問題を効率的に解くアルゴリズムである.

最適な状態列とは, 最もシンボル列の生成確率が高くなるような状態列のことである. あるモデル上において状態列とシンボル列が決定すれば, モデルがその状態列とシンボル列を生成する確率（尤度）は一意に求まる. 具体的には, シンボル列 $o_1 o_2 \cdots o_T$, 状態列 $q_1 q_2 \cdots q_T$ が与えられたときの尤度は

$$\pi_{q_1} b_{q_1}(o_1) \times a_{q_1 q_2} b_{q_2}(o_2) \times \cdots \times a_{q_{T-1} q_T} b_{q_T}(o_T)$$

と求まる. Viterbi アルゴリズムは, ある時刻 t でそれぞれの状態に到達する状態列のうち, 最も尤度の高い状態列のみを記憶していくことで最適な状態列を得る. Viterbi アルゴリズムは以下のように再帰的に尤度の最大値をとる計算を行う.

$$\delta_{t+1}(j) = \max_i (\delta_t(i) a_{ij}) b_j(o_{t+1})$$

この計算と同時に最大値を与える状態を記憶していけば, 最終的に最適な状態列を得ることができる.

2.3.3 モデルの算出

隠れマルコフモデルは 5 つのパラメータから成るが, このうち状態集合 Q とシンボル集合 Σ は事前に与えるパラメータである. 一方で状態遷移確率分布 A, シンボル出力確率分布 B, および初期状態確率分布 π は一般的に自明ではない. モデルの算出とは, これらの確率値を学習によって計算することである. モデルの算出を行うには, シンボル

列に内部状態をなんらかの方法（通常人手）によって与えたサンプルデータが必要になる．しかし, そのようなデータを利用できない場合でも, シンボル列のみの学習データによってパラメータを学習する Baum-Welch アルゴリズムによる学習が可能である．

Baum-Welch アルゴリズムは EM アルゴリズムの一種である．Baum-Welch アルゴリズムは, モデルが学習データとして与えられたシンボル列を生成する尤度が大きくなるようにパラメータの更新を繰り返すことで学習を行う．各繰り返しにおいて, 現在のパラメータによって各時刻における状態遷移の確率を求め, その期待値を最大化するようにパラメータを更新する．

$\overline{\pi}_i =$ 初期状態が状態 i の回数の期待値

$$\overline{a}_{ij} = \frac{\text{状態 } i \text{ から状態 } j \text{ へ遷移する回数の期待値}}{\text{状態 } i \text{ から遷移する回数の期待値}}$$

$$\overline{b}_i(k) = \frac{\text{状態 } i \text{ に滞在し記号 } k \text{ を出力する回数の期待値}}{\text{状態 } i \text{ に滞在する回数の期待値}}$$

この再推定式をパラメータが収束するまで繰り返し計算する．一般的に尤度は最大ではなく極大になるため, 初期パラメータの分布に依存して収束するパラメータは異なる場合がある．

2.4 トピック推定

ここでは本稿で提案する, 文書から事象系列を推定するアルゴリズムについて述べる．

2.4.1 HMM 手法の適用

我々は, トピックに隠れマルコフモデル（HMM）を適用してモデル化する．図 2.2 は, 事象を内部状態, 文章を出力シンボルに当てはめた HMM によるトピックの推定モデルである．前章で述べたように,HMM でモデル化することによって事象系列を Viterbi アルゴリズムに基づいて求めることができる．

しかし出力シンボルとして全ての単語を与えると, 効果的なトピック推定が困難になる．例えば, 東京都やアパートや事件といった単語は, トピックの状況の変化に対して意味を持っていない．そこで図中の左側に示すように, 文書中において特に状況の変化を表現する部分だけをモデルに反映させる．日本語の文章の場合, 特に状況の変化を表現する部分は各文章末の動詞である．このため図に例示されているように, 文書から各文章末の動詞部分のみを抽出し HMM のシンボルとして与える．

またトピックは, トピックの種類によって異なる事象で構成される．例えば, 殺人事件では事件発生や犯人逮捕といった事象があるが, 自殺事件では自殺発覚や理由発覚などといった異なる事象がある．つまり,HMM のパラメータの一つである状態の有限集

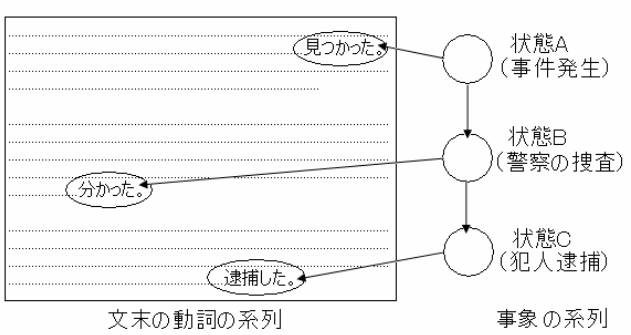


図 2.2: トピックの推定モデル

合 Q がトピックの種類によって異なる. このため, トピックの種類ごとに違う HMM を用意する必要がある.

ここで, 本節以降で使用する用語を定義する.

- 文書

  文書は, トピックを表現した文書という意味で用いる. 本稿では新聞記事の第一段落を日付順に連結し時系列に並べた文書のみを扱うが, ここでは特にその意味に限定するものではない.

- カテゴリ

  カテゴリは, トピックの種類という意味で用いる. 例えば, 殺人事件, 自殺事件, 汚職事件などがカテゴリとなる. 本研究では, カテゴリに依存して異なる HMM を用意する.

2.4.2 シンボル列の抽出

ある文書 d が与えられたとき, それに対応するシンボル列 $o_1 o_2 \cdots o_n$ を与える関数を考える. すなわち,

$$Symbol(d) = o_1 o_2 \cdots o_n$$

となるような関数 $Symbol$ を定義する.

文書は読点（。）で区切られた文章列とする. まず, それぞれの文章に対して形態素解析を行い, 単語列にする. 次に, 最後の形態素が過去を表す助動詞「た」でない文章を取り除く. これは, 死因の特定を急ぐ, 可能性もあるとみている, など状況の変化を伴わないシンボルを除去するためである. 最後の形態素が「た」である場合は, その直前に動詞があれば, その動詞をシンボル列に加える. この操作を文書 d の全ての文章に対して行う. これによって得られたシンボル列の末尾に, 終端を意味するシンボル「EOS」を加えたシンボル列を $Symbol(d)$ の値とする. このシンボル列のシンボルの順序は, 文書中の出現順序と一致していることを必ず保証する.

2.4.3 HMM モデルの学習

あるカテゴリ c に対応する HMM を M_c とする. M_c の学習用としてカテゴリ c の文書集合 $D_c = \{d_{c1}, d_{c2}, \cdots, d_{c|D_c|}\}$ が与えられたとき, M_c のパラメータを学習によって決定する方法を考える. M_c の状態集合 Q は任意の状態数 N_{M_c} 個の要素をもち, シンボル集合 Σ は全てのカテゴリの HMM で共通の集合とする.

まず,M_c の状態遷移確率分布 A, シンボル出力確率分布 B, 初期状態確率分布 π を乱数で初期化する. この M_c について,D_c のそれぞれの要素をシンボル列に変換して得られるシンボル列集合 L_c

$$L_c = \{Symbol(d_{c1}), \cdots, Symbol(d_{c|D_c|})\}$$

を学習データとして Baum-Welch アルゴリズムを実行し, パラメータを決定する.

なお, 最初に A,B,π を乱数で初期化するため, それぞれの状態がどのような事象の種類を意味しているかを事前に知ることができない. このため, この学習の後,HMM の確率分布を直接見ることで状態の解釈を後から加える.

2.4.4 トピックの推定

カテゴリの不明な文書 d が与えられたとき, d のトピックを推定する方法を考える.

まず,d によって与えられるシンボル列 $Symbol(d) = o_1 o_2 \cdots o_n$ に対して, 全てのカテゴリの HMM で状態列を推定する. いま, カテゴリ c の HMM,M_c が推定する状態列が $s_{c1} s_{c2} \cdots s_{cn}$ であるとする. このとき得た状態列とシンボル列の組を Mc が生成する確率 $P(o_1 o_2 \cdots o_n, s_{c1} s_{c2} \cdots s_{cn} | M_c)$ を最大にするような c が, 文書 d の所属するカテゴリである. すなわち,d の所属するカテゴリ c_d は

$$c_d = argmax_c P(o_1 o_2 \cdots o_n, s_{c1} s_{c2} \cdots s_{cn} | M_c)$$

であり, このカテゴリの HMM,M_{c_d} が推定した状態列が d のトピックとなる.

2.5 実験

ここでは, 提案アルゴリズムの評価実験について述べる. まず実験方法について述べ, 次に実験結果を示し, 最後に考察および提案アルゴリズムの評価を行う.

2.5.1 実験方法

我々は 3 つのカテゴリに分類した 256 件の文書を毎日新聞 2001 年,2002 年の 2 年分から人手で用意した. その内訳を表 4.1 に示す. それぞれのカテゴリで, 用意した文書のうち $\frac{2}{3}$ を学習用に, $\frac{1}{3}$ をテスト用に割り振る. カテゴリは次の 3 つである.

トピック	学習文書数	テスト文書数
単独犯事件	91	45
組織犯事件	35	17
汚職事件	46	22

表 2.1: 実験データ

- 単独犯事件：犯人が一人あるいは少数による殺人や強盗事件のトピック
- 組織犯事件：組織的な殺人や強盗事件のトピック
- 汚職事件：企業や政府などの要人による汚職事件のトピック

前節のアルゴリズムに従い, 学習文書を用いて HMM の学習を行い, テスト文書それぞれに対してトピックの推定を行う.

また今回, 全てのカテゴリで HMM の状態数 N_{M_c} を 5 で行った実験の結果を示す. 状態には事前にどのような事象の種類であるかという解釈を与えないため, 状態数は任意の値で学習を行わざるを得ない. このため, 予備実験によりカテゴリの分類率が最も高かった状態数 5 を用いる.

なお, 単語の切り分けおよび品詞の同定には日本語形態素解析ツール Chasen を用いる.

2.5.2 実験結果

それぞれのカテゴリについて,HMM のパラメータを学習した結果得られたモデルの構造を示す. ここで, モデルの構造（トポロジー）とは, 得られた HMM のパラメータが表現している状態間の関係を意味する. また, ここでは特に, 状態が出力するシンボルの分布も含める. この構造を見ることで, 状態を事象として解釈する.

図 2.3 は, 単独犯事件の HMM の構造である. 円は状態を表し, 矢印は遷移確率の大きい状態遷移を確率値と共に示している. また, 円の近くにはその状態が出力するシンボルのうち, 出力確率の大きいものを列挙してある.

これらの確率分布から, それぞれの状態が意味している事象を解釈する. 例えば, 状態 1 は 110 番通報があった, 見つけた, 刺された, などのシンボルが出力される確率が高いことから, 事件発生の事象を意味している. また, 状態 4 は逮捕した, 緊急逮捕した, 現行犯逮捕した, などのシンボルが出力されることから, 犯人の逮捕の事象を意味している. このように状態を解釈した結果を, それぞれの状態について図中の出力シンボルの下に示してある. 解釈は主観的な判断に基づいて行うが, このカテゴリのモデルの構造は比較的解釈が容易である.

次に, 状態の遷移に着目して遷移確率の高い状態をたどっていく. 事件発生, 警察の対応, 犯人逮捕, 事件収束と事象の変化として自然な状態遷移が起こりやすくなってい

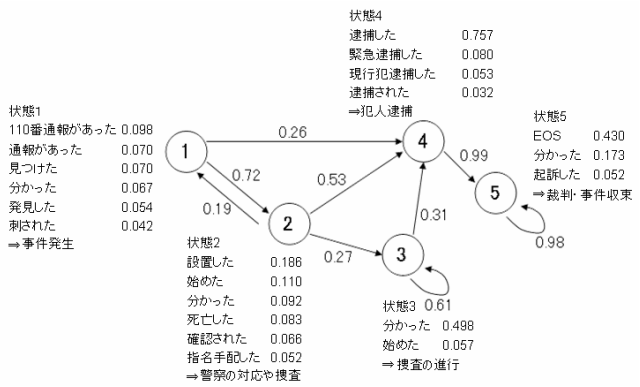


図 2.3: 単独犯事件の構造

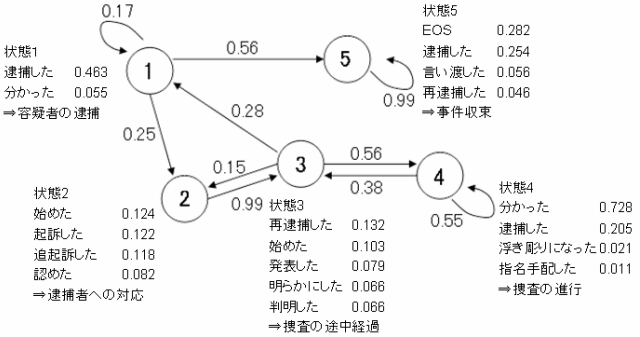


図 2.4: 組織犯事件の構造

る．また, 自分自身に遷移する確率の高い状態 3 を通るパスをもつ事件は, 捜査の長引く事件を表現している．状態遷移の点からも, このカテゴリのモデルの構造は解釈が容易になっている．

図 2.4 は, 組織犯事件の HMM の構造である．このモデルでは単独犯に比べ, 逮捕したのシンボルが複数の状態から出力される点で特徴的である．例えば, 捜査の進行と解釈した状態 4 からも高い確率で逮捕したのシンボルが出力されている．これは, 組織犯事件と分類したトピックには逮捕された犯人が複数いるような事件が多いため, 逮捕もまた捜査の進行の一部であると解釈する．

図 2.5 は, 汚職事件の HMM の構造である．このモデルは, 他のカテゴリに比べても複雑な構造である．例えば, 図中には捜査の進行と解釈した状態が二つある．これらの状態は出力シンボルの分布が似ており, 解釈を分けることができない．また, 状態遷移も複雑であり, 特徴的なパターンの発見が困難である．

表 2.2 に示すカテゴリの不明な文書に対して, トピックの推定を行った結果を一例として示す．

この文書から, シンボル列を抽出するアルゴリズムによって得られたシンボル列を

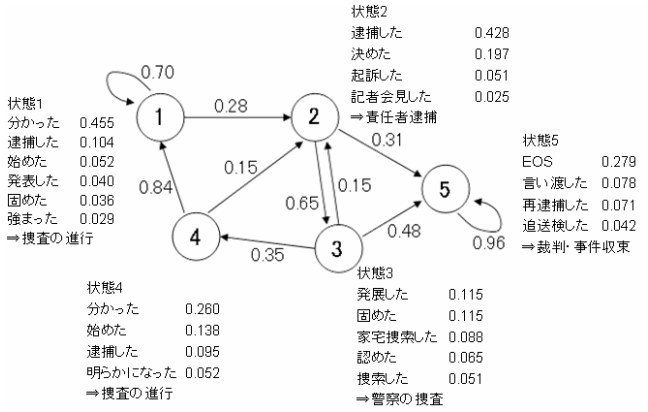


図 2.5: 汚職事件の構造

２３日午前２時３５分ごろ,... 署員が女性の焼死体を発見した。 ... 殺人事件と断定, 三島署に捜査本部を設置した。 ... ... , ○○さん（当時１９歳）が焼殺された事件で,... 疑いを強め, 事情聴取を始めた。 ... , ○○さん（当時１９歳）が殺害された事件で,... , △△容疑者（３０）を逮捕・監禁, 強盗などの疑いで逮捕した。 ... ... , △△容疑者（３０）が３０日,... ○○さん殺害を認める供述を始めた。 ... ... , 三島署捜査本部は１３日,... △△容疑者（３０）を殺人容疑で再逮捕した。 △△容疑者は容疑を認めている。

表 2.2: トピックを推定する文書

四角で囲んで示している．このシンボル列に対して推定された事象系列を表 2.3 に示す．なお, 表中で表記されている事象は, 図 2.3〜図 2.5 の図中に示した状態の解釈に対応する．

単独犯モデルが推定した事象は, 文書の解読から, 納得のいく結果になっている．特に, 逮捕したのシンボルが出現した後の事象が事件収束と推定されている．これは単独犯モデルが, トピックを犯人が一人の事件に限定しているためである．一方組織犯モデルでは, 逮捕したのシンボルが捜査の進行と推定されている．これは組織犯モデルが, トピックを犯人が複数いる事件に限定しているためである．

表の最後に示した尤度は, それぞれのモデルがこの状態列とシンボル列の組を生成する確率 $P(o_1o_2\cdots o_n, s_{c1}s_{c2}\cdots s_{cn}|M_c)$ である．この $argmax_c$ をこの文書のカテゴリと推定するため, この文書は単独犯事件のカテゴリに分類する．

テスト文書のトピックを分類した結果を表3.3 に示す．表にはカテゴリ別に正解率を

シンボル列	単独犯モデル	組織犯モデル	汚職モデル
発見した	事件発生	容疑者の逮捕	責任者逮捕
設置した	警察の対応や経過	逮捕者への対応	捜査の進行
始めた	捜査の進行	捜査の途中経過	捜査の進行
逮捕した	犯人逮捕	捜査の進行	捜査の進行
始めた	事件収束	捜査の途中経過	捜査の進行
再逮捕した	事件収束	容疑者の逮捕	事件収束
EOS	事件収束	事件収束	事件収束
尤度	6.25×10^{-9}	2.79×10^{-10}	5.87×10^{-18}

表 2.3: 推定された事象

トピック	正解数	テスト文書数	正解率 (%)
単独犯事件	34	45	75.6
組織犯事件	9	17	52.9
汚職事件	10	22	45.5
合計	53	84	63.1

表 2.4: 分類結果

示している.

3 つのクラスへの分類であるため,63.1% の正解率は評価できる数値である. 特に単独犯事件の分類率が 75.6% と高い. 一方, 汚職事件の分類率は 45.5% と最も低い結果である.

2.5.3 考察・評価

実験結果の考察と, 提案アルゴリズムの評価を行う.

まず実験結果から言えることは, 単独犯事件に関する結果がよいことである. モデルの構造の解釈が容易であり, 分類精度が特に高い. 逆に, 汚職事件はモデルの構造の解釈が難しく, 分類率が最も低い. この原因として, 汚職事件のトピックの事象系列が一定のパターンに従うことが少ないことがある. HMM は Baum-Welch アルゴリズムによって, 学習シンボル列の尤度を大きくするようにパラメータを収束させる. これによって得られたモデルは, 当然, 学習データと同じパターンのシンボル列の尤度を大きくする. もし他の汚職事件のシンボル列の尤度が小さいならば, それは学習したシンボル列とは異なったパターンのシンボル列である. モデルの構造の解釈が難しい原因も同様に, 学習シンボル列集合に多くのパターンが混在しているためである. ただし,HMM で近似したためにパターンが発見できなかった可能性はある. しかしながらこのモデルで, 単

独犯事件のパターンはよく表現できている. これより, 少なくとも,HMM でトピックを表現できるケースが存在すると言える.

また, どのカテゴリにおいても出現するシンボルが類似していることにも着目する. 今回用意した文書は, どれも事件に関するものであり, 逮捕した, 起訴した, 分かったなど類似したシンボルが出現する. このことから, 提案アルゴリズムは出現したシンボルの種類に依存してトピックを分類しているのではない. 提案アルゴリズムが考慮しているのはシンボルの出現順序であり, 従来の文書分類とは大きく異なっている.

2.6 結論

本研究では, トピックが事象の系列であるというアイデアに基づいて, トピックを確率過程としてモデル化する手法を提案した. また, 従来の文書分類とは異なり, 順序を考慮した分類を行えることを実験によって確かめた.

本研究では, 完結した事象系列をカテゴリ分類の対象とした. しかし, 未完結系列について本手法を適用することにより, 事件途中での予測が可能である. この推測から「今後の展開」に沿って捜査情報・手法の提示や対象の絞込みが行えるものとなろう.

本研究では, 教師有り学習として HMM モデルを予め構築し, いずれかのモデルに分類する方法をとったが, ニュースストリームからの逐次学習など, モデルの構築とモデル推定を同時に行わせることにより, 過去に例を見ない事件への適用も可能となる.

第3章　共起語を利用した事象系列に基づくトピック推定

本稿では, 文書のトピック推定の手法を提案する. 文書分類では一般的に, 文書を単語の出現頻度ベクトルなどでモデル化することが多いが, 本研究では, 文書として一連の事件を報じた新聞記事の系列を用いることで, 系列構造によるモデル化を行う. また, 事件は事象の系列であると考え, 文書に出現する動詞およびそれを特徴づける共起語の集合を事象に対応させ, 確率過程の手法に基づいて事象系列の尤度を与える方法を提案する. この方法によるトピック推定のアルゴリズムを示し, 実験によりその有用性を確認する.

3.1　前書き

近年, 計算機上で利用可能な文書データの増加に伴い, より高度な文書処理技術が必要とされている. この現状を背景にして, 文書分類技術に関する研究が盛んに行われている. 計算機による文書分類技術は, 一般的に文書データを出現単語ベクトルにモデル化し, 文書の類似度をベクトル間のコサイン尺度などで定義することによって, 自動的な分類を実現する. しかしベクトルモデルの欠点として, 単語の順序の情報を直接扱うことができないという問題がある. 文書が表現している内容によっては, 文章の順序が時間的な順序を表現しているなど, 順序が内容そのものに強く関係していることがある. 例えば, 新聞記事のように事件や出来事を客観的に記述した文書の場合, 内容の理解には順序を考慮する必要がある. このような文書を内容に基づいて分類するためには, 系列構造による文書のモデル化を行うことが望ましい. 本研究では, 特定の事件に関する新聞記事の系列を, ひとつの「事件」を記述した文書と考え, そのような文書を対象にしたトピック分類を目的とする.

トピックを扱う代表的なアプローチの一つに, *Topic Detection and Tracking*（TDT）がある [1]. TDT では, トピックは事象（event）によって特徴付けられる. 事象とは, 位置的, 時間的に特定の, 個々の発生した事実を意味する. TDT の Event tracking タスクでは, ある事象に関して述べている文書を逐次的に分類する [5]. Makkonen らは，出来事を分岐する事象の系列と考え，日付のある文書集合から出来事を発見する手法を提案している [8]. これらの研究では, 文書の類似度に基づいてトピックを発見する.

本研究では系列情報を反映したトピック分類の手法を提案する. 事象系列を考慮し

た分類手法は過去にあまり積極的な提案はなされていない. これは，決定木やSVM, 自己組織化マップ（SOM）, 単純ベイズ [9] といった従来の分類手法では, ベクトルの分類に問題を帰着させるため，系列情報を反映させることが容易ではないことによる. 本稿では, トピックを事象系列のクラスと考え, 確率過程に基づく文書分類を提案する.

我々はこれまでに, 動詞を事象の特徴と考え文書をモデル化し, トピック推定を行う手法を提案している [20]. ここでは, 文末の動詞が文書の事象としての要約になっていることを利用して, 隠れマルコフモデルのシンボルに動詞を対応させ, 事象系列のトピック推定を行う. 実験では, 隠れマルコフモデルが文書の分類器としてうまく働くことを示している.

しかし, 表現する事象によっては, 動詞単独では事象の記述として十分でないことがある. 例えば,「捜査を始めた」と「事情聴取を始めた」など, 同じ動詞でも共起する格によって意味が大きく異なることがある. 本稿では, 動詞の共起語を事象の特徴量として利用し, より文書の内容を反映させたトピック分類の手法を提案する.

関連研究として, 文書を確率過程を用いてモデル化する手法に,Mulbregt らがある [11]. ここでは文書を話題の系列と考え, 系列の隠れ状態を推定することで文書のトピックセグメンテーションを行っている. 彼らはセグメンテーションを目的としているため, 系列そのものには意味を対応させず, 話題の遷移確率は話題の境界に対するペナルティーとしてのみ作用している.

文書の構造を系列でモデル化して推定する研究に,Barzilay らがある [2]. ここでは, 地震などを報じる文書には報じる内容の順序に特徴があることを利用して, 確率過程モデルを用いて文書の構造を推定する. 彼らは. 文章の bigram 分布を内容に対応させ, あらかじめクラスタリングした文章を学習データとして分布のパラメタを推定する. この手法は単語分布に依存するため, 文書の表現のパターンを扱うのに適しているが, この方法をそのままトピックのパターンに適用することはできない.

確率過程による文書のモデル化を談話構造の解析に応用する研究に, 柴田らがある [4]. ここでは日本語による料理番組のナレーションを対象として, 用言の格フレームをシンボルとみなした隠れマルコフモデルを用いている. 実験では, 動詞に着目することで, 事象の遷移をうまく捉えられることを示している.

2 章では事象系列に基づくトピック分類について述べる. 3 章では本研究で用いる確率過程モデルについて説明する. 4 章で本稿で提案するトピック推定のアルゴリズムの説明を行う. 5 章で実験結果を示し,6 章で結びとする.

3.2 事象系列によるトピックの分類

本稿では,「事象」は位置的, 時間的に特定される個々の事柄, 発生した事実を意味する. 「トピック」は, 一連の事件やテーマに関する事象系列のクラスを意味する.

２つの事件が似ているかどうかは, 類似性の定義によって異なる. 本研究では事象の系列が似ている事件を「類似する」とする. 例えば, 東京で起きた強盗事件と広島で起

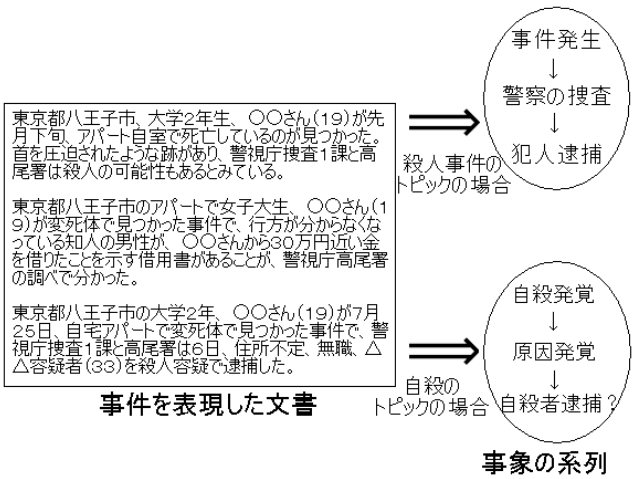


図 3.1: 文書からのトピック推定

きた強盗事件は, 場所も犯人も盗品も違う. しかし, どちらも「盗まれた」,「指名手配された」,「犯人が逮捕された」等のように, 事象の系列が類似しているため, ここで両事件は「類似する」とする.

図 3.1 は, トピック推定の例である. 「殺人事件」のトピックを表現している文書集合に対し, 新聞記事の第一段落を日付順に連結して時系列に並べる. この例は, ある殺人事件に関する新聞記事を連結したものである.

図の右側には, この文書でたどっている「事象の系列」を示す. 個々の事象は, この事象系列が何のトピックであるかによって解釈が異なる. 記事内容は, ある人物の死亡の発見, 警察による犯人の手がかりの発覚, 容疑者が逮捕されるという事象を表し, これらの事象が,「殺人事件」という特性を述べていると考えることができる. 他方, この文書がある人物の自殺に関するトピックならば, 推定される事象の種類として, 自殺の発覚のあとに自殺の原因の特定が続くであろう. 殺人事件の最後では,「逮捕」に関する事象が通常生じる. 自殺事件の場合でも, 逮捕された容疑者が獄中で自殺を図ることが考えられるが, すでに自殺発覚の事象があるため, 発生順序が不自然である. 即ち, 殺人事件トピックは自殺トピックとは通常両立しない系列を有している.

このように, 事象の発生順序にはトピックごとに特徴があると考えられるため, 本稿では, 事象の遷移に着目することで事件のトピック分類が可能であることを論じる.

3.3 確率過程モデル

ここでは, トピック推定に用いる確率過程モデルを定義する. 本研究では, 確率過程モデルとして確率過程オートマトンを用いる. 確率過程オートマトンは, 確率的な状態遷移と各状態からの確率的な出力をもつオートマトンである. 本稿では, 状態は単純マ

ルコフモデルに基づいて遷移する.

モデルの枠組みは隠れマルコフモデル [3] と同じであるが, 隠れマルコフモデルでは状態は観測できず, 主に状態の推定問題に適用される. これに対し, 本稿で用いるモデルでは状態は観測可能であり, 系列の確率を求めることのみを目的とする.

3.3.1 モデルの定義

与えられた状態の集合 Q と, 状態から出力されるシンボルの集合 V について, 確率過程オートマトンを定義する. 本稿では, 状態には動詞が対応し, シンボルには動詞の共起語が対応する. 例えば,「警察が捜査を始めた」という文章からは, 状態として「始める」, シンボルとして「警察」「捜査」を得る.

本稿で用いる確率過程オートマトンは, 以下に示す 5 つのパラメタ (Q, Σ, A, B, π) で定義する.

(1) $Q = \{q_1, \cdots, q_N\}$：状態の有限集合

Q は状態の集合である. それぞれの要素は動詞に対応する.

(2) $\Sigma = \{o_1, \cdots, o_M\}$：出力の有限集合

Σ はシンボルの集合 V の冪集合である. すなわち,$\Sigma \subseteq 2^V$ であり,o_t は V の部分集合である. 各状態は 1 つの Σ の要素を出力する. これは,1 つの動詞が任意の数の共起語を伴って出現することに対応する.

(3) $A = \{a_{ij}, i, j = 1, \cdots, N\}$：状態遷移確率分布

a_{ij} は状態 q_i から状態 q_j への遷移確率であり, $a_{i1} + \cdots + a_{iN} = 1.0$ である. 動詞 i が観測された後, 次に観測される動詞の確率分布に対応する.

(4) $B = \{b_i(o_t), i = 1, \cdots, N, t = 1, \cdots, M\}$：シンボル出力確率分布

$b_i(o_t)$ は状態 q_i が o_t を出力する確率であり, $b_i(o_1) + \cdots + b_i(o_M) = 1.0$ である. 動詞 i が観測されたとき, 同時に観測される共起語の確率分布に対応する. ひとつの動詞が複数の共起語を伴って出現しているとき,$b_i(o_t)$ はそれぞれの共起語 $o_{t,k}$ の出力確率の積として求める. すなわち, シンボル集合 $o_t = \{o_{t,1}, o_{t,2}, \cdots, o_{t,|t|}\}$ が状態 i から出力される確率は,

$$b_i(\{o_{t,1}, o_{t,2}, \cdots, o_{t,|t|}\}) = \prod_k b_i(o_{t,k})$$

とする.

(5) $\pi = \{\pi_i, i = 1, \cdots, N\}$：初期状態確率分布

π_i は状態 q_i が初期状態である確率である. 文書で最初に観測される動詞が i である確率に対応する.

本稿では, 確率行列 A は単純マルコフ過程に基づく状態遷移確率に対応する. これは, 次の状態の確率は現在の状態にのみ依存することを意味する. 例えば,「見つかる, 始める, 逮捕する」という状態列において「逮捕する」が観測される確率は, 前の状態「始める」にのみ依存して決定される. 同様に, シンボルの出力確率は現在の状態にのみ依存する. 状態とシンボル集合は, 語として文書から観測できる.

3.3.2 モデルの算出

本稿の確率過程モデルは 5 つのパラメタ (Q, Σ, A, B, π) から成るが, Q および Σ は事前に与える. ここではモデルの算出として, 状態遷移確率分布 A, シンボル出力確率分布 B, および初期状態確率分布 π を学習によって計算する方法について述べる. 本研究では, モデルの算出に教師あり学習を用いる [10]. 各トピックに人手で分類した学習データを用いて, トピックにおける事象系列パターンをモデルに反映させる.

まず, 学習データ D から状態とシンボルの系列を抽出し, 状態遷移回数, シンボルの出力回数, 初期状態の回数をカウントする. それぞれの回数の相対頻度を, 確率値として用いる. すなわち,D_i を最初の動詞が i である文書の集合,V_{ij} を動詞 i から j への遷移回数, V_i を $\Sigma_j V_{ij}$,$W_i k$ を動詞 i が語 o_k と共起して出現した回数, W_i を $\Sigma_k W_{ik}$ とすると, それぞれのパラメタは以下の推定式で求める.

$$\pi_i = \frac{|D_i|}{|D|}, \quad a_{ij} = \frac{V_{ij}}{V_i}, \quad b_i(k) = \frac{W_{ik}}{W_i}$$

モデルのパラメタが決定しているとき, 系列の確率を求めることができる. 次節に示すように, 系列は動詞列に対応する状態列 $q = \{q_1, q_2, \cdots, q_T\}$ と, 共起語集合列に対応するシンボル集合列 $o = \{o_1, o_2, \cdots, o_T\}$ の組から成る. 例えば,< (始める, { 警察, 捜査 }), (逮捕する, { 殺人, 未遂, 容疑 }) > のような系列を考える. モデル M が q, o を出力する確率 $P(q, o|M)$ は,

$$P(q, o|M) = \pi_{q_1} b_{q_1}(o_1) \prod_{t=1}^{T-1} a_{q_t q_{t+1}} b_{q_{t+1}}(o_{t+1})$$

と求める.

3.4 トピック推定

ここでは本稿で提案する, 文書のトピック推定アルゴリズムについて述べる. 本研究では, 尤度原理に基づいてトピック推定を行う. それぞれのトピックに対応する確率過程モデル $\mathcal{M}_1, ..., \mathcal{M}_{\mathcal{L}}$ においてテストデータの尤度を求め, 尤度を最大にするトピック m を文書のトピックと推定する. 本章では, 推定アルゴリズムの詳細を示す.

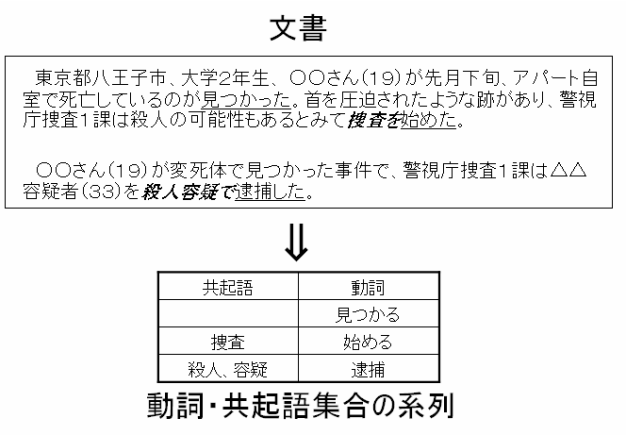


図 3.2: 動詞と共起語の抽出

3.4.1 確率過程モデルの適用

本研究では, 各トピックに確率過程モデルを対応させ, 尤度原理を適用することでトピック推定を行う. 本稿では，文書を一つのトピックに関する記事系列とする. それぞれの新聞記事の第一段落には最も重要な内容が含まれていると考え，各記事の第一段落を時系列順に連結したものを文書として扱う.

事件は, トピックの内容に依存した事象の種類から成る. 例えば, 殺人事件トピックでは「容疑者の逮捕」,「被害者の発見」といった事象が考えられるが, 自殺事件トピックでは「遺書の発見」など異なる事象が生じる.

しかし, 入力文書に出現する全ての語をモデルに反映させると特徴的な事象系列パターンの推定が困難になる. このため本研究では, 図 3.2 に示すように, その時点で起きた事象を表現する部分として文末の動詞とそれに関連する共起語のみを用いる. これは, 日本語の文章において, 状況の変化を表現するために各文章末の動詞を用いることが多いという特徴による. しかし日本語の場合, 動詞に関連する共起語は主格や目的格に関わらず, 動詞よりも前の, どの位置にも出現する可能性がある. これは, 助詞によって格を表現するという日本語の特徴による. 本研究では,「Chasen」[6] による形態素解析を行い, 共起語辞書を用いて動詞と共起語を抽出する.

本研究では, 共起語の抽出に EDR 電子化辞書[1]の日本語共起辞書を用いる. この辞書は, コーパスの解析により抽出された語句同士の係り関係を列挙したものであり, 各語句に対して品詞や意味の情報が付与されている. 本稿では, この日本語共起辞書から名詞句から動詞句への係り関係のみを取り出したものを「共起辞書」と呼び, 動詞の共起語の抽出に用いる. 共起辞書の一部を表 3.1 に示す.

本研究では動詞を状態に, 共起語をシンボルに対応させ, 文書を系列で形式化する. また, トピックごとに学習文書を用意し, トピックに対応する確率過程モデルを学習によっ

[1]情報通信研究機構 EDR 電子化辞書
http://www2.nict.go.jp/r/r312/EDR/J_index.html

名詞句	動詞
警察が	逮捕
犯人を	逮捕
犯人は	逮捕
電話を	設置
本部を	設置
少年が	走る
痛みが	走る
アプリケーションが	走る

表 3.1: 共起辞書

て得る.

3.4.2 状態およびシンボルの抽出

ある文書 D が与えられたとき, 対応する系列 $< q_1 q_2 \cdots q_n, o_1 o_2 \cdots o_n >$ を与える関数 g を考える. ここで,q_i は i 番目の状態であり,o_i は i 番目のシンボル集合である.

文書は読点（。）で区切られた文章列とする. まず, それぞれの文章に対して形態素解析を行い, 単語列にする. 次に, 最後の形態素が過去を表す助動詞「た」でない文章を取り除く. これは,「死因の特定を急ぐ」,「可能性もあるとみている」, など状況の変化を伴わない記述を除去するためである. 最後の形態素が「た」である場合は, その直前に動詞があれば, その動詞を取り出し,v_t とする.

次に,v_t と共起関係にある名詞句を抽出する. v_t と同じ文章に含まれる全ての名詞句について, 一番後ろに現れている名詞以降の形態素を文字列として取り出す. この文字列について共起辞書を参照し,v_t との共起レコードが存在するか調べる. レコードが存在する場合, その名詞句に含まれる名詞を全て共起語として抽出し,n_t とする. 共起語がひとつも抽出されず,n_t が空集合である場合は, 共起語が無いことを意味するシンボル「NONE」を n_t に加える.

例えば,「太郎を殺人容疑で逮捕した」という文章からは, 動詞として「逮捕」, 名詞句として「太郎を」,「殺人容疑で」を抽出する. 「殺人容疑で」という名詞句は,「殺人」「容疑」「で」の 3 形態素から成るため, 一番後ろに現れている名詞である「容疑」以降の文字列「容疑で」を取り出す. 共起語辞書を参照し,「容疑で, 逮捕」のレコードが存在する場合, 「殺人」,「容疑」を共起語として抽出する. 共起辞書にレコードが存在しない場合, その名詞句を破棄する.

この操作を文書 D の全ての文章に対して行い, 得られた v_t の系列を状態列 q, n_t の系列をシンボル列 o とする. さらに,q の末尾に終端を意味する状態「EOS」を加え, それに対応するシンボルとして「NONE」を o の末尾に加える. q_t および o_t の順序は, 文書

中の出現順序に一致させる. 文書 D について関数 g は

$$g(D) =< (q_1, q_2, \cdots, q_T, EOS), (o_1, o_2, \cdots, o_T, NONE) >$$

となる.

3.4.3 モデルの学習およびトピックの推定

ここでは, 学習によってモデルのパラメタを算出し, トピック推定を行う手法を示す.

トピック c に対応するモデルを $\mathcal{M}_c$ とし, $\mathcal{M}_c$ の学習用としてトピック c の文書集合 D_c が与えられているとする. D_c に含まれる全ての文書について $g(D)$ を求め, 状態遷移回数, シンボル出力回数, 初期状態の回数をカウントする. この頻度から,3.2 節で示した推定式を用いてモデル $\mathcal{M}_c$ のパラメタを決定する.

本研究では尤度原理を適用し, テスト文書 D のトピックを推定する. $g(D) =< q, o >$ について, 全てのトピック c に対応するモデル $\mathcal{M}_c$ が $< q, o >$ を生成する確率 $P(q, o|\mathcal{M}_c)$ を求める. この確率は 3.2 節で示した, 遷移確率とシンボル出力確率の積を求める式を用いて計算する.

文書 d のトピックは,$P(q, o|\mathcal{M}_c)$ の値を最大にするようなトピックと推定する. すなわち, 推定する文書 d のトピック c_d は,

$$c_d = argmax_c P(q, o|\mathcal{M}_c)$$

とする.

3.5 実験

3.5.1 実験方法

本稿では 3 つのトピックに分類した 256 件の文書を, 毎日新聞 2001 年,2002 年の 2 年分から人手で用意する. その内訳を表 4.1 に示す.

ここで扱うトピックは次の 3 つである.

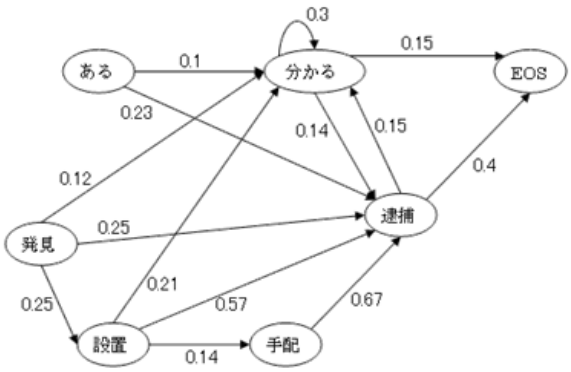
- 単独犯事件：犯人が一人あるいは少数による殺人や強盗事件のトピック
- 組織犯事件：組織的な殺人や強盗事件のトピック
- 汚職事件：企業や政府などの要人による汚職事件のトピック

それぞれのトピックで, 用意した文書の一部を学習用, 残りをテスト用とする. 提案アルゴリズムに従い, 学習文書を用いてモデルの学習を行い, テスト文書それぞれに対してトピックの推定を行う. トピック推定の結果と人手による分類との一致率で, 提案アルゴリズムの評価を行う.

単語の切り分けおよび品詞の同定には日本語形態素解析ツール「Chasen」[6] を用いる. 共起辞書には,EDR 電子化辞書の「日本語共起辞書」を用いる.

トピック	学習文書数	テスト文書数
単独犯事件	91	45
組織犯事件	35	17
汚職事件	46	22

表 3.2: 実験データ



動詞	共起語
ある	通報、110番
発見	遺体、一室、路上
設置	捜査、本部、県警、署
手配	容疑、殺人、強盗、傷害
分かる	こと、調べ、供述、可能性、疑い
逮捕	容疑、疑い、殺人、死体、逮捕、県警、強盗、傷害

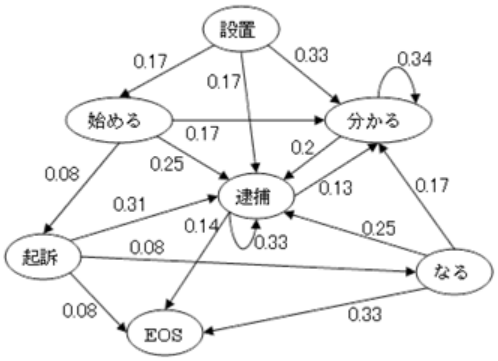
図 3.3: 単独犯事件モデルの構造

3.5.2 実験結果

まず, それぞれのトピックについて学習アルゴリズムで得られた確率過程モデルの構造を示す．ここで, モデルの構造（トポロジー）とは, モデルの状態間の関係および状態が出力するシンボルの分布を意味する．この構造を見ることで, トピックの事象系列の特徴をモデルが表現していることを確かめる．

図 3.3 に, 単独犯事件のモデルの構造の一部を示す．円は状態を表し, 矢印は遷移確率の大きい状態遷移を確率値と共に示している．円の中には, 状態が対応する動詞が示してある．また図の下には, それぞれの状態から出力されるシンボルのうち, 確率の大きいものを列挙してある．

この構造から, 単独犯事件モデルの特徴を考察する．まず, 出力される確率の高い共起語から,「ある」は「通報があった」,「発見」は「遺体が発見された」という文脈で出現することが多い．これらの動詞は, 初期状態である確率が高く, 他の状態からの遷移確率は低い．これは単独犯事件モデルにおける事件発生の特徴を表している．また,



動詞	共起語
設置	捜査、本部、県警、特別、対策
起訴	容疑者、地検、罪、事件、殺人
始める	捜査、調査、容疑、死体、逮捕、殺人、本部
なる	浮き彫り、事件、捜査、可能性、明確、段階
分かる	こと、調べ、供述、実行、疑い、可能性、目撃
逮捕	容疑、疑い、殺人、死体、逮捕、保険金、違反

図 3.4: 組織犯事件モデルの構造

「逮捕」から「EOS」への遷移確率が高いことは, 単独犯事件の特性から, 一人の犯人が逮捕されることで事件が解決する特徴を表していると言える.

図 3.4 は, 組織犯事件のモデルの構造である. このモデルでは,「逮捕」から再び「逮捕」に遷移する確率が高い. これは, 組織犯事件トピックには犯人が複数いるような事件が多いため, 複数の逮捕が続けて起こるという特徴を表していると考える.

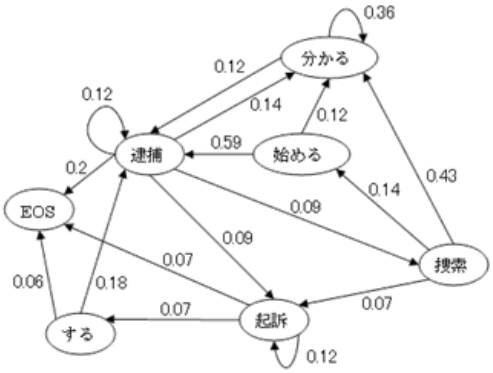
図 3.5 は, 汚職事件のモデルの構造である. このモデルは他のトピックに比べ, 共起語に特徴がある. 例えば,「逮捕」には「贈賄」「収賄」「詐欺」,「始める」には「家宅捜索」,「する」には「懲戒免職」といった単語が共起しやすい. これらの単語は汚職事件の特徴を表している.

テスト文書のトピック分類を行った結果を表 3.3 に示す. 表にはトピック別に正解率を示している. 全トピックの合計で 64.3% の正解率を得た. 単独犯事件の分類率が 71.1% と最も高く, 組織犯事件の分類率は 47.1% と最も低い.

3.5.3 考察・評価

実験結果の考察と, 提案アルゴリズムの評価を行う.

第一に, 本実験の正解率が示すように, 単独犯事件に関する結果がよい. 文書を詳しく見ると, 単独犯事件の特徴として犯人が一人であるため, 逮捕に関する事象で事件が終わることが多い. これは単独犯事件モデルの特徴と一致しており, 分類率が高い一因と考えられる.



動詞	共起語
始める	取り調べ、家宅、捜索、容疑、聴取、疑い
捜索	自宅、特捜、地検
起訴	事件、地検、容疑者、地裁、罪
する	処分、こと、懲戒、免職、事件、証言、電話
分かる	こと、調べ、話す、供述、疑い
逮捕	容疑、疑い、贈賄、背任、収賄、違反、詐欺、妨害

図 3.5: 汚職事件モデルの構造

トピック	正解数	テスト文書数	正解率 (%)
単独犯事件	32	45	71.1
組織犯事件	8	17	47.1
汚職事件	14	22	63.6
合計	54	84	64.3

表 3.3: 分類結果

先行研究 [20] との比較では, 汚職事件の分類率が先行研究で 45.5% であったのに対し, 本実験では 63.6% と大きく改善されている. これは文書の特徴量として共起語を新たに用いたことによる. 実際, 学習したモデルの構造を見ると, 汚職事件で頻出する共起語は他の 2 つのトピックと比べ特徴的である.

一方, 組織犯事件は分類率が最も低い. これは, 出現する動詞および共起語が単独犯事件と類似しており, さらに単独犯事件に比べ, 組織犯事件には多様な事象系列パターンが存在するためであると考えられる. しかしながら, 文書に現れる語が類似していても 47.1% とある程度の分類が可能であることから, 提案アルゴリズムが出現する語に依存しておらず, 従来の文書分類とは本質的に異なることが言える.

学習したモデルの共起語を見ると, 動詞と共起語の関係が不明な場合がある. 例えば, 「設置」の共起語として「本部」が高頻度で現れているが, 「本部が設置した」,「本部を設置した」など, 複数の関係で共起することが考えられる. しかし, 日本語文章では同

一の意味でも異なる格助詞を伴うことがあり, また受身の表現では同一の格助詞でも異なる関係を意味するなど, 表層の情報から動詞との関係を与えることは容易ではない.

また, 本手法では固有名詞を共起語としてモデルに反映することが難しい. これは, 発生頻度が少ないため共起辞書に含まれていない可能性が高いことと, シンボルとしての発生頻度が少ないためシンボル出力確率が大きくならないことによる. 例えば, 国外逃亡の事象では「アメリカに逃亡した」,「中国に逃亡した」など固有名詞が共起語に現れることで本手法によるパターンの発見が困難になる.

3.6 結論

本研究では, 事象系列の分類という目的に基づいて, 確率過程を用いたモデル化によるトピック分類の手法を提案した. また, 動詞の共起語を用いることにより, 動詞だけでは表現できない事象の記述をモデルに反映させる手法を示した. 従来の文書分類とは異なり, 順序を考慮した分類を行えることを実験によって確かめた.

本研究では, 完結した事象系列をトピック分類の対象とした. しかし, 未完結系列について本手法を適用することにより, 事件途中での予測が可能である. この推測から「今後の展開」に沿って捜査情報・手法の提示や対象の絞込みが行えるものとなろう.

本研究では, 教師有り学習としてモデルを予め構築し, いずれかのモデルに分類する方法をとったが, ニュースストリームからの逐次学習など, モデルの構築とモデル推定を同時に行わせることにより, 過去に例を見ない事件への適用も可能となる.

第4章　差分型HMMを用いたデータストリームにおける時系列データ予測

本研究では時系列データ予測の新しい手法を示す. ここでは各時刻の観測データを事象として解釈し, 事象の遷移に基づいてデータの予測を行う. 本稿では事象を隠れ状態に対応させた隠れマルコフモデル (HMM) によるモデル化を行い, 状態からの出力確率分布を予測値の分布として用いる. また, データストリーム環境においてはEMアルゴリズムによるHMMのパラメタ学習を行うことが困難であるため, 差分型Baum-Welchアルゴリズムによるオンライン学習手法を適用し, 実験により本手法の有用性を検証する.

4.1　前書き

近年, 計算機資源の充実化に伴い, より知的な情報処理を計算機で行う研究が盛んになっている。特に, 計算機を用いることで膨大な量のデータを扱えるようになったことを背景にして, データベースからの知識発見やデータ予測の技術が注目されている. 計算機による時系列データ予測とは, 時刻を伴ってリアルタイムに観測されるデータについて, 現在までの観測データを用いて未来に観測されるデータを自動的に推定する技術である.

時系列データ予測は, リアルタイムに大量の情報を扱う場合に有用な技術である. 例えば, 商品の在庫状況の管理を計算機によって行っているシステムにおいて, 時系列データ予測が可能となれば, 翌日の売り上げを予測することで効果的な商品の仕入れが自動的に行うことができる. また, 大量のデータを利用して客観的にデータ予測を行うことで, 企業や政府の意思決定を補助する効果も期待される.

時系列データ予測は, これまでに多くの手法が提案されている [15]. 指数平滑法 (Exponential Smoothing) は, 過去のデータの重みつきの平均値に基づいて予測を行うヒューリスティックなデータ予測手法である. 重みは, 観測から経過した時刻に対して指数的に減少する. このとき, 時刻 t での重みつきの平均値は, ひとつ前の時刻 $t-1$ の重みつき平均値を用いて集約的に求めることができる.

Holt-Winters 法は指数平滑法の一種であり, 重みつき平均値の増減の傾向を考慮して

予測を行う [14]. ここでは, 時刻 t の平滑化した観測値を $\tilde{y}_t$, その変化量を F_t として, 各時刻で以下の式を用いて $\tilde{y}_t$ および F_t を更新する.

$$\tilde{y}_t = \lambda_1 y_t + (1 - \lambda_1)(\tilde{y}_{t-1} + F_{t-1})$$

$$F_t = \lambda_2(\tilde{y}_t - \tilde{y}_{t-1}) + (1 - \lambda_2)F_{t-1}$$

ただし,λ_1 および λ_2 は平滑化パラメタと呼ばれる 0.0 から 1.0 までの値であり, λ が大きいほど過去のデータの重みは急速に減少する. Holt-Winters 法による時刻 $t + h$ の予測は,$\tilde{y}_t$ を用いて

$$\tilde{y}_{t+h|t} = \tilde{y}_t + hF_t$$

として求める. これは, 未来のデータが現在の増加量の傾向を維持することを意味する.

しかし, これらの手法は観測値の平均または差分が時間に対して大きく変化しないことを前提とする. イベント会場に出入りした人数の時系列データや, ある地点の風速の時系列データなどは, 前の時刻の観測値と直接関係せず, 催し物や台風の接近といった「事象」に基づいて観測値が決まる. このような時系列データに対しては, 指数平滑法で効果的な予測が可能であるとは考えにくい.

Hassan ら [16] は, 隠れ状態をもつ確率過程モデルである隠れマルコフモデルを用いて株価の予測を行う. ここでは過去の観測データから, 隠れマルコフモデルの学習により頻出する推移のパターンを抽出する. 学習したモデルを用いて, 最新の時刻の観測データにどの隠れ状態が対応するかを確率的に求め, 次の時刻に観測されるデータの尤度を推定することで予測を行う. 隠れ状態の確率を求めることは, 観測されるデータに対して, 最新の観測データが過去のどの推移パターンと類似するかを推定することに対応する. この手法では, 予測値は最新の時刻の観測に直接依存せず, 隠れ状態の遷移に基づいて決定する. このことから, 隠れた事象を含む時系列データに対しても効果的な予測が期待できる.

しかし,Hassan らは EM アルゴリズムを用いて, オフラインに隠れマルコフモデルの学習を行うため, 予測の最中に未知の系列パターンが出現した場合に, 新しいパターンをモデルに反映させることができない. このことは, 特にデータストリーム環境において大きな問題となる [17][18]. データストリームではデータは永続的に発生し, また動的にデータの分布が変化するため, 学習データに含まれるデータのパターンはデータストリーム全体で発生するパターンのごく一部である. このため, モデルの学習を予測を行いながら（オンラインに）行う必要がある.

隠れマルコフモデルの学習を差分的に行う研究に,Stenger らがある [19]. 彼らはラベルなし学習データによる隠れマルコフモデルの学習手法である Baum-Welch アルゴリズムを拡張した, 差分型 Baum-Welch アルゴリズム (Incremental Baum-Welch) を提案している. 差分型 Baum-Welch アルゴリズムでは, 新規のデータに対して推定した隠れ状態の尤度に基づいて, その尤度が最大になるようにパラメタの再計算を行う. この手

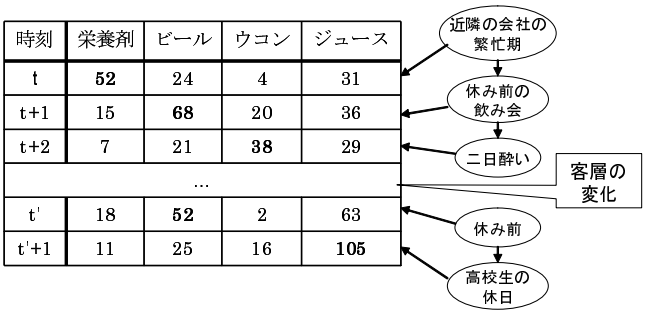


図 4.1: 状態遷移に基づく時系列のモデル化

法は, 更新前のパラメタを用いて集約的にパラメタの更新値を求めるため, 過去の観測データを記憶することなく, かつ高速に計算できる. Stenger らは, 差分型 Baum-Welch アルゴリズムを用いて動画中の物体検出を行っている.

本稿では, 差分型 Baum-Welch アルゴリズムを用いた時系列データ予測の手法を提案する. これにより, 隠れマルコフモデルによる適応的な時系列データ予測が可能になることを論じる.

2 章では状態遷移に基づくデータ予測について述べる. 3 章では本研究で用いる隠れマルコフモデルのパラメタの動的な推定法を説明し, 本稿で提案するデータ予測のアルゴリズムについて述べる. 4 章で実験結果を示し,5 章で結びとする.

4.2 状態遷移に基づくデータ予測

図 4.1 に, 本研究で提案する時系列データ予測のモデルを示す. 図中の時系列データは, 量販店の商品の売り上げ履歴である. ここでは, 時刻 t に栄養剤がよく売れ, 翌日 (時刻 $t+1$) にはビール, その翌日 (時刻 $t+2$) にはウコンがよく売れている.

本研究では, 時系列データにおいて観測されるデータは, 観測できない隠れた状態に依存すると考える. 例えば, 図 4.1 の時刻 t ではビールの売り上げが少なく, 栄養剤がよく売れていることから, 来店する客は仕事で忙しい人が多いと解釈する. 時刻 $t+1$ では, 栄養剤よりもビールやおつまみがよく売れていることから, 翌日が休日などの理由でお酒を飲む人が多い日と解釈する. 同様に時刻 $t+2$ ではウコンがよく売れることから, 二日酔いの人が多い日と解釈する.

これらの状態の対応づけから, 仕事が忙しい人が多い日の翌日はお酒を飲む人が多い日に遷移し, お酒を飲む人が多い日は二日酔いの人が多い日に遷移する. これらのパターンが過去の観測データ系列の中に頻繁に出現していれば, 時刻 t の時点で時刻 $t+1$ や時刻 $t+2$ の売り上げの予測が可能である.

ただし, 実際にはこれらの状態の解釈は行われる必要はない. 隠れ状態の遷移に基づく時系列データのパターンをなんらかのアルゴリズムによって獲得できれば, 状態の意味が解釈されなくてもデータ予測は可能である. 本研究では, 隠れマルコフモデルによ

り隠れ状態を明示的にモデル化し, 状態の意味を事前に与えることなく, 観測したデータ系列に基づいてモデルのパラメタを決定する.

図 4.1 では時間の経過に伴い, 近くに高校ができるなどの原因により客層が変化し, これにより状態の意味や遷移パターンが変化する. 例えば, 高校生の客が増えることで休日のジュースや菓子の売り上げが増加するといった変化が考えられる. このような変化を反映して予測を行うためには, 常に新しいパターンの存在を考慮する必要がある.

隠れマルコフモデルにおいて, 新しい観測データから未知のパターンをモデルに反映することは容易ではない. これは, モデルが隠れ状態を含むため, 観測値から直接最尤のパラメタを推定できないことによる. 非観測の状態を含む確率モデルでは, 一般に EM アルゴリズムによりパラメタの推定を行う. EM アルゴリズムは, まず現在のパラメタを用いて隠れ状態の確率を求め, その確率に基づいて新たにパラメタを更新する. そして再度更新したパラメタを用いて同じ操作を行い, パラメタが収束するまでこれを繰り返す. 隠れマルコフモデルのラベルなし学習を行うアルゴリズムとして,EM アルゴリズムの一種である Baum-Welch アルゴリズムが存在する. しかし,Baum-Welch アルゴリズムは繰り返し計算を全てのデータに対して行うため, 一般に長大なデータになるほど実行時間は膨大となる. このため, 新しい観測値が利用可能になるたびに Baum-Welch アルゴリズムによる再計算を行うのは困難である.

特にデータストリーム環境では, 新しい観測に対して差分的にパラメタを更新することで動的な新規パターンの反映を実現する必要がある. これは, データ量が膨大であるために過去のデータを保存できないことと, 全データの再計算を実行することが計算量の観点から困難であることによる. 差分的なパラメタ更新とは, なんらかの集約値を用いて過去のデータを少ない情報量で表現し, これを用いてパラメタを更新することを指す. 差分型 Baum-Welch アルゴリズムは, 隠れマルコフモデルの各パラメタを過去のデータの集約値として用いることにより, 差分的なパラメタ更新を実現する.

本研究では差分型 Baum-Welch アルゴリズムを適用することにより, 新規パターンに適応的な時系列データ予測を実現することを目指す.

4.3 差分型 HMM を用いたデータ予測

4.3.1 隠れマルコフモデル

隠れマルコフモデル (Hidden Markov Model, HMM) は, 確率的に遷移する隠れ状態をもつオートマトンである. 隠れ状態は, 各時刻で単純マルコフ過程に従って遷移する. 状態は観測できないが, 各時刻の状態は確率的に観測値を出力する. また, 観測値はその時刻の状態にのみ依存する.

隠れマルコフモデルは次の 5 つのパラメタによって定義される [3].

(1) $Q = q_1, ..., q_N$：状態の有限集合.

(2) $\Sigma = o_1, ..., o_M$：観測値の集合.

(3) $A = a_{ij}$：状態遷移確率分布. a_{ij} は状態 q_i から状態 q_j への遷移確率である.

(4) $B = b_i(o_t)$：観測値の出力確率分布. $b_i(o_t)$ は状態 q_i で観測値 o_t を出力する確率である.

(5) $\pi = \pi_i$：初期状態確率分布. π_i は状態 q_i が初期状態である確率である.

本稿では, 観測値として多次元の連続値ベクトルを考える. ここでは出力確率分布 $b_i(o_t)$ は各次元で独立に正規分布に従う. このため本モデルは, 状態毎, および観測値の次元毎に正規分布のパラメタとして平均と分散をもつ.

(6) μ_{ik}：観測値平均. μ_{ik} は状態 q_i において k 次元目の要素の確率分布を与える正規分布の平均である.

(7) σ_{ik}：観測値分散. σ_{ik} は状態 q_i において k 次元目の要素の確率分布を与える正規分布の分散である.

時刻 t の観測ベクトル o_t の次元 k の要素 o_{tk} が出力される確率は, 状態 q_i に滞在しているとき, 平均 μ_{ik}, 分散 σ_{ik} の正規分布に従う.

$$b_i(o_{tk}) = \frac{1}{\sqrt{2\pi}\sigma_{ik}} e^{-\frac{(o_{tk}-\mu_{ik})}{2\sigma_{ik}}}$$

観測ベクトル o_t の出力確率は, それぞれの要素 o_{tk} の出力確率の積として求める. すなわち,$o_t = (o_{t1}, o_{t2}, \cdots, o_{tD})$ が状態 i から出力される確率は,

$$b_i(o_{t1}, o_{t2}, \cdots, o_{tD}) = \prod_{k=1}^{D} b_i(o_{tk})$$

である.

与えられた学習データについて最適な状態遷移確率分布 A, 観測値平均 μ, 観測値分散 σ を推定する際には, ラベルなしデータを用いて準最適なパラメタを求める Baum-Welch アルゴリズムを用いる. 本研究では状態の意味はあらかじめ分かっておらず, 観測データのパターンを自動的に学習することが望ましい. Baum-Welch アルゴリズムは EM アルゴリズムの一種であり, 与えられた学習データの尤度が大きくなるようにパラメタを繰り返し更新し, 準最適なパラメタに収束させる.

4.3.2 差分型 Baum-Welch アルゴリズム

時間の経過等により新たな観測データを得たとき, モデルのパラメタを更新する手法について述べる. 差分型 Baum-Welch(Incremental Baum-Welch) アルゴリズムは, まず現在のモデルを用いて, 新規のデータに対応する時刻 T における状態 i の尤度 $\gamma_T(i)$ を計算する. モデルは現在の時刻 $T-1$ における状態 i の滞在確率 $\gamma_{T-1}(i)$, および現在ま

での全ての時刻の状態 i の滞在確率の和 $\Sigma_{t=1}^{T-1}\gamma(i)$ を保持しているとする. 時刻 T で状態 i から状態 j に遷移する確率 $\gamma_T(i,j)$ は,

$$\gamma_T(i,j) = \gamma_{T-1}(i)\frac{a_{ij}b_j(o_T)}{\Sigma_j a_{ij}b_j(o_T)}$$

で求まる. 時刻 T での状態 i の滞在確率は $\gamma_T(i) = \Sigma_j\gamma_T(i,j)$ である. ここから, 各パラメタは次のように更新される [19].

$$a'_{ij} = \frac{\Sigma_{t=1}^{T}\gamma_t(i,j)}{\Sigma_{t=1}^{T}\gamma_t(i)} = \frac{\Sigma_{t=1}^{T-1}\gamma_t(i)a_{ij} + \gamma_T(i,j)}{\Sigma_{t=1}^{T}\gamma_t(i)}$$

$$\mu'_i = \frac{\Sigma_{t=1}^{T}\gamma_t(i)o_t}{\Sigma_{t=1}^{T}\gamma_t(i)} = \frac{\Sigma_{t=1}^{T-1}\gamma_t(i)\mu_i + \gamma_T(i)o_T}{\Sigma_{t=1}^{T}\gamma_t(i)}$$

$$\sigma'_i = \frac{\Sigma_{t=1}^{T}\gamma_t(i)(o_t-\mu_i)^2}{\Sigma_{t=1}^{T}\gamma_t(i)} = \frac{\Sigma_{t=1}^{T-1}\gamma_t(i)\sigma_i + \gamma_T(i)(o_T-\mu_i)^2}{\Sigma_{t=1}^{T}\gamma_t(i)}$$

差分 Baum-Welch アルゴリズムは, モデルのパラメタを再推定式の集約値として用いることにより, 過去の観測データを保持することなく, 高速に更新を行うことができる.

4.3.3 差分型 HMM によるデータ予測

本稿で提案する, 差分型 HMM を用いたデータ予測の手法を示す.

差分 Baum-Welch アルゴリズムは EM アルゴリズムではないため, 乱数で初期化したモデルのパラメタの収束は遅い. ここでは, 最初から差分 Baum-Welch アルゴリズムによる学習を適用せず, まず既知である観測系列を用いて Baum-Welch アルゴリズムにより初期モデルを生成する. これは, パラメタの収束が極端に遅くなるのを防ぐためである.

まず, 状態数 N の HMM のパラメタとして, 状態遷移確率分布, 観測値平均を乱数で決定し, 観測値分散を全て 1.0 とする. このモデルについて, 既知である観測系列を用いて Baum-Welch アルゴリズムを実行する. パラメタが収束したら, 時刻 T までの各時刻 t で状態 i に滞在する確率 $\gamma_t(i)$ を計算し, その和 $\Sigma_{t=1}^{T}\gamma_t(i)$ および最終時刻 T での滞在確率 $\gamma_T(i)$ を保存する.

得られた初期モデルを用いて, 次の時刻 $T+1$ の観測値の予測を行う. $T+1$ の観測値が未知であるとき, 時刻 $T+1$ で状態 i に滞在する尤度 $\gamma_{T+1}(i)$ は,

$$\gamma_{T+1}(i) = \Sigma_{j=1}^{N}\gamma_T(j)a_{ji}$$

で与えられる. 状態 i に滞在するとき, 時刻 $T+1$ の最も尤もらしい観測値は, $\mu_i = (\mu_{i1}, \mu_{i2}, ..., \mu_{iD})$ である. 以上のことから, 時刻 $T+1$ の観測値の期待値は,

$$\Sigma_{j=1}^{N}\gamma_{T+1}(i)\mu_i$$

であり, この値を時刻 $T+1$ の予測値とする.

時刻が経過し, 時刻 $T+1$ の観測値 o_{T+1} を観測したとき, 差分型 Baum-Welch アルゴリズムに従ってパラメタの更新を行う. このとき, 時刻 $T+1$ の状態の滞在確率は, 更新されたパラメタを用いて求める. すなわち,

$$\gamma_{T+1}(i) = \Sigma_{j=1}^{N}\gamma_T(j)a'_{ji}b'_i(o_{T+1})$$

として更新し, 次の時刻の予測に用いる.

4.4 実験

4.4.1 実験方法

本稿では「気象データひまわり」の 1996 年から 2000 年の 5 年分,1828 日分のデータベースから, 「日最大瞬間風速 (m/s)」「日照時間 (hour)」「日降水量 (mm)」の 3 項目の予測を行う. ここでは, 東京, 大阪, 福岡の 3 観測地点についてのデータを実験に用いる. 表 4.1 に, 東京の観測データ系列を示す.

1828 日のデータのうち, 学習データとして系列の始めから 10 %から 50 %のデータを用いて残りの観測データの予測を行い、結果を比較する. また, 予測アルゴリズムとして, 差分 Baum-Welch を使用しない HMM による予測法 (HMM), 差分 Baum-Welch を使用した HMM による予測法 (Inc HMM), およびベースラインとして Holt-Winters の Exponential Smoothing 法 (H-W) の 3 つで予測を行い, 精度を比較する. 予測精度の尺度は二乗誤差平均 (Mean Square Error; MSE) を用いる. 時刻 t の予測値を p_t, 実際の観測値を x_t とすると,MSE は次のように定義される.

$MSE = \frac{1}{T}\sum_{t=1}^{T}(p_t - x_t)^2$

また, 全実験データの 90 %について, 差分 HMM と通常の HMM を用いて予測を行い, 実行時間を比較する. 実行時間は 10 回の試行の平均により評価する.

本稿では HMM の状態数は 15 として実験を行う.

4.4.2 実験結果

表 4.2 に, 各アルゴリズムによる福岡の瞬間最大風速, 日照時間, 降水量の予測の MSE を示す. ここに示したのは, 学習データとして全実験データの 20 %にあたる 365 日分の観測データを用いたときの結果である. 表中の平均は, 最大瞬間風速の MSE, 日照時間の MSE, 降水量の MSE を合計して属性数 3 で割った値である.

時刻	最大瞬間風速	日照時間	降水量
1996 年 6 月 29 日	16.1	4.8	0.0
1996 年 6 月 30 日	11.0	1.4	14.0
1996 年 7 月 1 日	8.4	2.6	0.5
1996 年 7 月 2 日	7.6	0.8	0.0
1996 年 7 月 3 日	7.9	0.9	6.5
1996 年 7 月 4 日	10.5	0.1	0.0
1996 年 7 月 5 日	16.8	0.7	3.5

表 4.1: 東京の実験データ

手法	最大風速	日照時間	降水量	平均
HMM	13.29	18.79	196.82	76.30
差分型 HMM	13.14	15.89	190.99	73.34
Holt-Winters	21.33	23.35	316.43	120.37

表 4.2: 各アルゴリズムによる福岡の観測の予測誤差 (MSE)

この条件においては, 各属性とも差分型 Baum-Welch を使用した HMM による予測で最も MSE が小さくなる. Holt-Winters 法と比較すると, 差分型 HMM の MSE は平均で約 39 %の改善がみられる. 差分型 Baum-Welch を使用しない HMM との比較では, 平均で約 4 %, 最大では日照時間の予測において約 15 %の MSE の改善となる.

図 4.2,4.3,4.4 に, それぞれ東京, 大阪, 福岡の MSE の, 各次元の平均値を示す. グラフの横軸は学習に用いた系列の長さが全体の実験データの内で占める割合である. これらの結果から, 本実験データについて HMM を用いた手法は Holt-Winters 法よりも小さい誤差で予測が可能であるといえる. また, 通常の HMM 手法と差分型 HMM 手法では予測誤差に劇的な違いはないが, 特に学習データが少ないときの予測精度で差分型 HMM 手法による精度の改善がみられる.

表 4.3 に, 予測の実行時間を示す. これは全実験データの 90 %にあたる 1646 件の観測データの予測に要した時間である. 1 観測あたりの実行時間は, 実行時間を観測データ数 1646 で割った値である. パラメタ更新を実行する分, 差分型 HMM 手法は実行時間がかかるが, その差は 1 観測あたり 0.16(ms) とほとんど差がない. 差分 Baum-Welch アルゴリズムは, 動的なパラメタ更新を極めて高速に行うことができると言える.

4.4.3 考察

実験結果の考察と, 提案アルゴリズムの評価を行う.

図 4.2,4.3,4.4 から, Holt-Winters 法の予測精度は HMM を用いた手法に比べて非常に

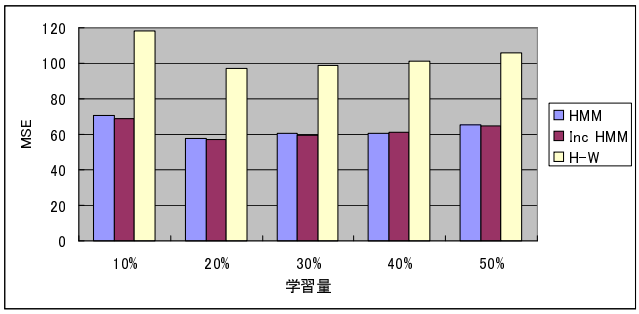


図 4.2: 東京の平均予測精度

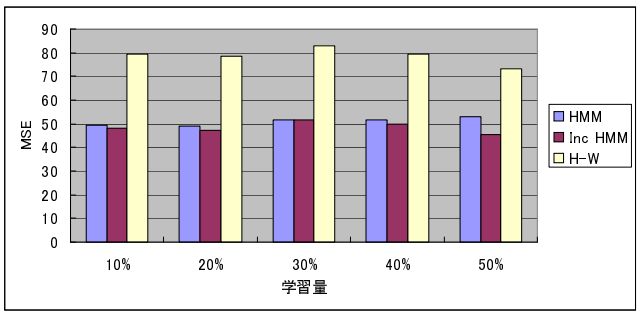


図 4.3: 大阪の平均予測精度

悪い．これは, 観測値の時間方向の差分値に基づいて予測を行うことが原因である．このため, 本実験データのように一日ごとに観測値の差分が大きく変化するデータに対しては効果的な予測ができない．また, 各次元の観測値を独立に予測を行うため, 予測に利用できる情報が少ないことも予測精度を悪化させる要因である．

差分型 HMM と通常の HMM の比較では, 学習量の全データに対する割合が小さいほど差分型 HMM 手法の方が通常の HMM の予測精度を上回る傾向がみられる．これは, 少ない学習データ中には出現しなかった系列パターンを差分型 Baum-Welch によって動的にモデルに反映させているためである．データストリーム環境においては, 一般に学習データとして利用可能な系列は予測を行うデータに比べ非常に少ないため, この結果は提案手法がデータストリーム環境において効果的であることを示している．

大阪のデータの予測においては, 学習データの割合が大きいにも関わらず差分型 HMM 手法が通常の HMM の予測精度を大きく上回っている．これは主に, 学習量を 50 %にしたことで通常 HMM 手法の降水量の予測精度が悪化したことによる．学習データを増加させたにも関わらず予測精度が悪化したのは, 学習に用いたデータの特性に起因する．実験データの解析から, 大阪では 98 年の降水量が他の年に比べて少なく, また雨が続かない特徴的な系列を有している．学習量を 40 %から 50 %に増加させたことで 98 年の系列を学習データに取り込んだため, 翌年以降の予測精度を悪化させたと考えられる．

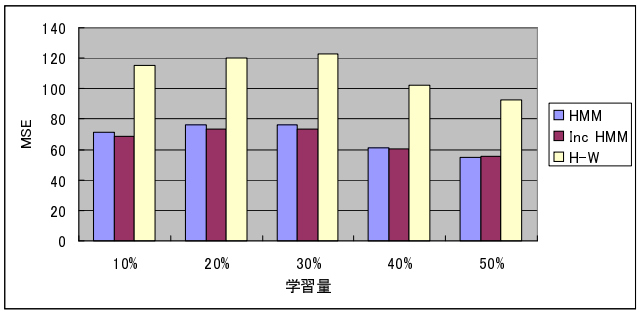


図 4.4: 福岡の平均予測精度

手法	実行時間 (ms)	1 観測あたりの実行時間 (ms)
HMM	623.3	0.38
差分型 HMM	892.1	0.54

表 4.3: 予測の実行時間

一方, 差分型 Baum-Welch を行う HMM 手法では, 通常 HMM 手法と同じモデルから予測を開始していながら, 翌年以降の傾向をモデルに反映させることで予測精度の悪化を防いでいる. このことから, 提案手法が学習データの局所的な特性に依存することなく, 未来のデータについて適応的に予測を行うことが可能であるといえる.

また, 実行時間の観点から, 差分型 Baum-Welch アルゴリズムは通常の HMM 手法と比較してわずかな計算で実行可能である. 本手法は実行速度の面でも, データストリーム上での予測に十分適用可能であるといえる.

4.5 結論

本研究では, 時系列データの予測を行う新しい手法として, 差分型 HMM を用いた手法を提案した. 本手法により, 学習データが少ない環境においても適応的にオンライン学習を行い, 高速に予測を行えることを実験によって確かめた.

本研究では, モデルの状態数はあらかじめ与えられているものとした. しかし, 差分パラメタ更新に加えて状態数を増減することにより, 過去に例を見ないパターンの学習をより素早く行えるものとなろう.

本研究では観測値は数値データとしたが, 本手法は確率過程モデルを用いているため非数値データへの適用も可能である. 例えばニュースストリームに適用することにより, 事象の予測を行うといった応用も考えられる.

第5章　結論

本研究では, 時系列的特徴に基づくパターン発見手法として, 隠れ状態を含む確率過程モデルで時系列データをモデル化する手法の提案を行った. 本手法が観測値の関係に直接依存せず, 観測値を隠れ状態として解釈することで頑強な時系列特徴の発見を実現できることを示した. また, 状態として解釈することで, 単語や単語集合, 多次元ベクトルデータといった様々な観測に対してパターンの学習を行えることから, 本手法は汎用性のある有用な手法であるといえる.

まず, 新聞記事のトピック推定では, 文書で記述された一連の事件を時系列的特徴に基づいて分類する新しい手法を提案した. 本手法により, 時系列データとしての文書について, 内容の遷移パターンに基づいて適切に分類を行えることを実験によって確かめた. ここでは, 隠れマルコフモデルが, 観測データである文書中の動詞の表記のゆれを吸収し, 事象として解釈しなおす仕組みとして機能している. Baum-Welch アルゴリズムによる教師なし学習を行うことで, この解釈のプロセスを時系列特徴からの動詞のクラスタリングとしてみなせることが, 本実験によって明らかとなった.

文書の共起語を用いた詳細な特徴の抽出では, 共起辞書を用いることで重要語に関連した特徴的なシンボルを得る手法を提案した. 実験により, 本手法を用いることで文書から事象に強く関連する特徴を抽出でき, 時系列パターンを適切に捉えた事件の分類ができることを確かめた. ここでは, 重要語に意味的に関連する単語を, 同一文章内に多数存在する冗長な語の中から共起辞書という知識を用いることで効果的に抽出できることを示している. 動詞の遷移が出来事の特徴パターンとなりうるばかりでなく, その共起語を考慮することで動詞の意味が限定され, より効果的な特徴パターンを得られることを示した.

本手法を用いることで, 文書から適切な要約部分を抽出することにより, 文書を時系列データとして扱うことが可能である. 本手法は, 時系列データとしての特徴を持つウェブ上の文書に対して適用することで, そのウェブページが有する時系列特徴を捉えることも可能となるであろう.

差分型 HMM を用いた時系列データ予測では, 隠れマルコフモデルによって時系列データのモデル化を行うことで, 事象に基づいた時系列データの予測を行うことができることを示した. また, 差分型 Baum-Welch アルゴリズムを適用することにより, データの分布が動的に変化するデータストリーム環境においても, 新しい時系列パターンに適応的かつ高速な予測が可能であることを示した. 従来の HMM を用いた時系列予測に関する研究では, オフライン学習データが十分にある環境での実験しか行われていなかったが, 本研究では学習データが十分でない場合においても効果的な予測が行える手

法を示した. また, 局所的な学習データの特徴から予測精度が低下する問題があることからも, オンライン学習を用いた予測手法が必要であるといえる.

本手法はベクトルデータを観測値とした時系列データを予測の対象としたが, 隠れマルコフモデルの出力確率分布の定義を変えることによって, 様々な種類のデータの予測に適用できる. 例えば, 上述の文書のトピック推定で用いた出力確率分布の定義を用いれば, 文書内容の予測が可能になる. 今後, 様々な種類のデータに対して時系列パターンを抽出し, より有用な知識の獲得を行うアルゴリズムの研究が求められる.

謝辞

本研究を遂行するにあたり，日頃より適切な御指導，御鞭撻をいただいた，法政大学工学部情報電気電子工学科 三浦孝夫教授に深く御礼申し上げます．

並びに，産能大学経営情報学科 塩谷勇教授にも多くの有益なご助言をいただきました．深く感謝いたします．

データ工学研究室の先輩方，同輩，後輩たちにも，本研究の遂行にあたって数多くの助言と快適で能率的な研究室環境を整えていただきました．御礼申し上げます．

修士論文として私の研究をまとめることができたのも，多くの皆様方の御支援，御協力の賜物であります．この場をお借りしまして，厚く御礼申し上げます．

最後に，今までの学生生活を支えてくださった私の両親に深く感謝したいと思います．

参考文献

[1] Allan, J., Carbonell,J., Doddington, G., Yamron,J. and Yang,Y.: “Topic Detection and Tracking Pilot Study: Final Report”, proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[2] Regina Barzilay and Lillian Lee: “Catching the Drift: Probabilistic Content Models, with Applications to Generation and Summarization”, In Proceedings of the NAACL/HLT, pp. 113-120, 2004.

[3] 北 研二: “確率的言語モデル”, 東京大学出版会, 1999.

[4] 柴田 知秀, 黒橋 禎夫: “隠れマルコフモデルによるトピックの遷移を捉えた談話構造解析”, 言語処理学会 第 11 回年次大会, 2005.

[5] Yiming Yang, Tom Ault, Thomas Pierce, Charles W. Lattimer: “Improving Text Categorization Methods for Event Tracking”, In Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval, 2000.

[6] Asahara, M. and Matsumoto, Y.: Extended Models and Tools for High Performance Part-of-Speech Tagger, COLING, 2000

[7] D. M. Blei and P. J. Moreno.: “Topic segmentation with an aspect hidden Markov model”, In Int. Conf. Research and Dev. Inf. Retrieval, pp. 343–348, New York, 2001.

[8] Makkonen, J.: Investigations on Event Evolution in TDT, In Proceedings of HLT-NAACL 2003 Student Workshop, May 2003, Edmonton, Canada, pp. 43-48.

[9] Manning, C.D. and Schutze, H.: Foundations of Statistical Natural Language Processing, MIT Press, 1999

[10] Mitchell, T.: Machine Learning, McGrawHill Companies, 1997

[11] Mulbregt, P.van; Carp, I.; Gillick, L.; Lowe, S.; and Yamron, J. 1998. Text Segmentation and Topic Tracking on Broadcast News Via a Hidden Markov Model Approach. In Proceedings of the ICSLP’98, volume 6. 2519–2522.

[12] Cavalin P.R., Sabourin R., Suen C.Y. and Britto Jr., A.S.: Evaluation of Incremental Learning Algorithms for An HMM-Based Handwritten Isolated Digits Recognizer, The 11th International Conference on Frontiers in Handwriting Recognition (ICFHR 2008), Montreal, August 19-21, 2008.

[13] S. Duan and S. Babu: Processing forecasting queries. In Proc. of the 2007 Intl. Conf. on Very Large Data Bases, 2007.

[14] Sarah Gelper, Roland Fried, and Christophe Croux: Robust Forecasting with Exponential and Holt-Winters Smoothing, (June 2007)

[15] Jan G. de Gooijer, Rob J. Hyndman: "25 Years of IIF Time Series Forecasting: A Selective Review," Tinbergen Institute Discussion Papers 05-068/4, 2005.

[16] Md. Rafiul Hassan , Baikunth Nath: StockMarket Forecasting Using Hidden Markov Model: A New Approach, Proceedings of the 5th International Conference on Intelligent Systems Design and Applications, 2005.

[17] Jian, N. and Gruenwald, L.: Research Issues in Data Stream Association Rule Mining, SIGMOD Record 35-1, 2006, pp.14-19

[18] S.Muthukrishnan: Data streams: algorithms and applications, Proceedings of the fourteenth annual ACM-SIAM symposium on discrete algorithms (2003).

[19] B. Stenger and V. Ramesh and N. Paragios: Topology free Hidden Markov Models: Application to background modeling, In IEEE International Conference on Computer Vision, 2001.

[20] Kei Wakabayashi, Takao Miura: Identifying Event Sequences using Hidden Markov Model, 12th Intn'l Conf. on Applications of Natural Language to Information Systems (NLDB), Springer LNCS 4592, pp.84-95

[21] Kei Wakabayashi, Takao Miura: Topics Identification Based on Event Sequence Using Co-occurrence Words, 13th Intn'l Conf. on Applications of Natural Language to Information Systems (NLDB),2008