進化的トピックモデルの提案とその実現

横山　慎† 馬　強†

† 京都大学 大学院情報学研究科　〒 606–8501 京都府京都市左京区吉田本町
E-mail: †yokoyama@db.soc.i.kyoto-u.ac.jp, ††qiang@i.kyoto-u.ac.jp

あらまし　ニュースやマイクロブログなどの時系列テキストデータでは，時間経過とともにトピックが進化する．従来のトピックモデルは，その多くが一括学習であることから，時間経過に伴うトピック数などの変化に対応できず，トピックの新規発生や混合などの進化を効率よく捉えられない場合が多い．そのため，我々は時系列テキストデータについて，時間経過に伴う新しいトピックの発生や既存のトピックの混合を推定する新しいトピックモデルについて研究を行っている．本研究では，学習済みモデルでの予測精度の継時変化をもとにしたモデル更新時期の検知と，遺伝的アルゴリズム等での新しいトピック分布を作成することによるモデルの更新を繰り返すことで，新しい話題に対し新しいトピック分布を継続的に生成可能な進化的トピックモデルの実現を目指す．実験では，本手法の実現に向け perplexity，coherence に基づく時系列テキストデータに対するトピックの変化点検知に焦点を当てた検証・考察を行う．なお，本稿ではモデル更新時期の検知手法に焦点を当てた検証を行うため，新しいトピック分布の作成には上述の手法ではなく，便宜的に学習データとトピック数を更新した LDA を用いる．

キーワード　トピックモデル，確率過程，遺伝的アルゴリズム

1 はじめに

近年，インターネットの普及に伴い，数多くのニュースサイトやマイクロブログサービスが運営されている．Web 上でのニュース記事の配信やマイクロブログサービスでの投稿によって，膨大な量のテキストデータが日々新たに蓄積されて続けている．テキストデータの潜在意味解析に際し，もっともよく用いられる手法の一つにトピックモデルがある．トピックモデルは文書中の潜在的意味を解析する手法であり，代表的なものに Blei らの Latent Dirichlet Allocation (LDA) [1] がある．LDA は，文書が固有のトピック分布を持つことを仮定した文書生成モデルである．

従来の LDA モデルの多くでは，学習用文書データは一括で全て与えられ，パラメータの学習を行っている．学習に際して，学習用文書の順序の交換可能性を仮定したり，学習データを通じて一定なトピック数をパラメータとして与える必要がある．LDA の拡張として，Dynamic Topic Model(DTM) [2] や Topic Tracking Model(TTM) [3] といった時系列を考慮したトピックモデルや，Hierarchical Dirichlet Process(HDP) [4] などの適切なトピック数を自動推定するトピックモデルも提案されているが，これらにおいてもトピック数は継時変化せず一定であるのに対し，ニュースやマイクロブログサービスにおいては話題は常に移り変わる．時間経過によって全く新しい話題が現れたり，既存の話題であっても含有する単語割合が徐々に変化するため，これらのモデルは効率よく処理できない場合があると考えられる．

図 1 は，既存手法での適切な学習が困難と考えられる話題について，2 つの例文と生成されうるトピック分布を用いて説明したものである．例えば仮想通貨は．2017 年頃に出現した全く新しいトピックで，それまでのどのトピックにも類似しない．また，「山中先生」や「研究」といった単語は京都大学トピックで，「マラソン」「完走」といった単語はスポーツトピックで出現しやすい単語で，それぞれ新しい単語ではないものの，これらが一つの文脈内で語られることは未だなかった，というようなトピックの発生も考えられる．先のトピックモデルは，変化しないトピック数のもとでトピック分布の時系列変化をとらえたり，単にトピック数の自動推定を行うものであり，新しい話題に対して新しいトピック分布を生成し割り当てることは保

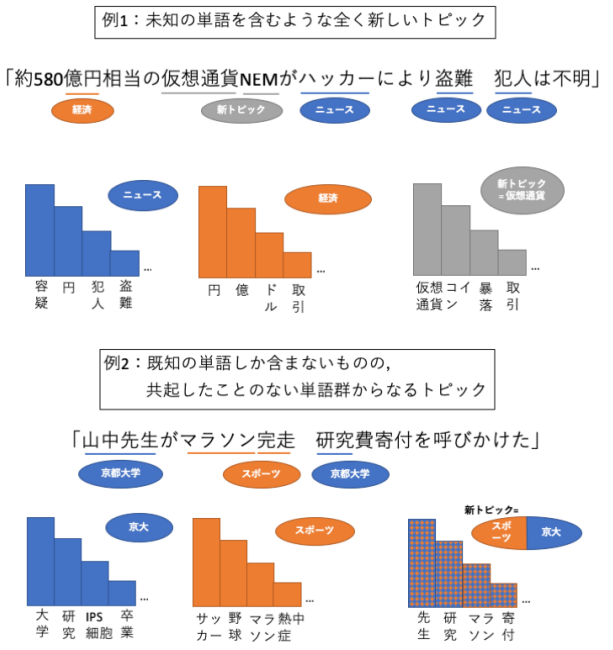

図 1　LDA での適切な学習が困難と考えられる話題の例

証しない．新しい話題に対して個別にトピック分布を与えることができれば，各トピックを明確に区別することを可能にし，文書内容を考慮したクラスタリングや推薦を行う際に有利になる．本研究ではこのような文書の集合についてトピックの変化をとらえ，トピック分布を生成する手法を提案する．

そのため，我々はこのようなニュース記事やマイクロブログの投稿といった時系列情報を持つテキストデータについて，時間経過に伴う新しいトピックの発生や既存のトピックの混合を推定する新しいトピックモデルについて研究を行っている．学習済みモデルでの予測精度の継時変化をもとに新しい話題の生起を検出し，遺伝的アルゴリズムによって対応するトピック分布を生成することで，新しい話題に対して新しいトピック分布が生成され割り当てられることを保証する仕組みを研究している．

本論文では，文書毎の perplexity の時間経過に伴う推移をもとにしたモデル更新時期の検知と，遺伝的アルゴリズム等で新しいトピック分布を作成することによるモデルの更新を繰り返し続けることで，新しい話題に対し新しいトピック分布の差分的な生成を実現する手法を提案する．

また，時系列テキストデータでは，出現する話題は逐次変化している．更に文書そのものの数も無限に増加しており，すべての文書を用いて一括学習を行うことは困難である．このため本研究では，最初に一括学習を行った後，以後の新たな文書を用いて学習済みモデルを逐次更新することで，各文書を一度しか学習に必要としない差分的な手法を提案する．これにより，一括学習のみによる再学習と比較して高速かつトピックの時間変化を考慮した学習が可能になる．

また，文書が増加するにつれ，話題の数も無限に増加する．新しいトピック分布を都度生成し続けると，その分メモリを圧迫し，継続的な学習が困難になる．そのため本研究では図 1 下部のような例に対しては新トピックを既存トピックの交叉により生成する．これにより，少ないメモリ量での継続的な学習が可能になる．

本論文の構成は次の通りである．第 2 節でトピックモデルやその時系列拡張，モデルの評価に関わる研究について述べた後，第 3 節で本論文で提案する進化的トピックモデルについて説明する．第 4 節では進化的トピックモデルの実現に向けて行なった実験についての説明と結果，考察を示す．第 5 節では今後の課題とともに本論文の結論を述べる．なお，本稿ではモデル更新時期の検知手法に焦点を当てた検証を行うため，新しいトピック分布の作成には上述の手法ではなく，便宜的に学習データとトピック数を更新した LDA を用いる．

2 関 連 研 究

2.1 トピックモデル

トピックモデルとは，確率モデルの一種であり，あるドキュメントが複数のトピックの混合として成り立っていると仮定した文書生成モデルである．文書はトピックの混合分布，トピックは単語の混合分布として表現される．代表的なトピックモデルとして頻繁に用いられる LDA では，入力データとして文書の bag-of-words 表現を，潜在変数として単語の持つ潜在トピックを導入している．図 2 に LDA のグラフィカルモデルを示す．

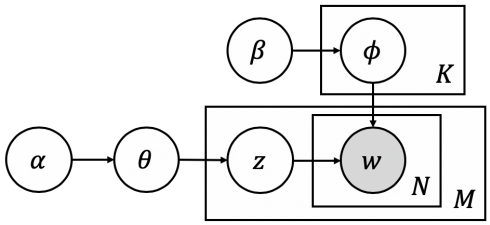

図 2　LDA のグラフィカルモデル

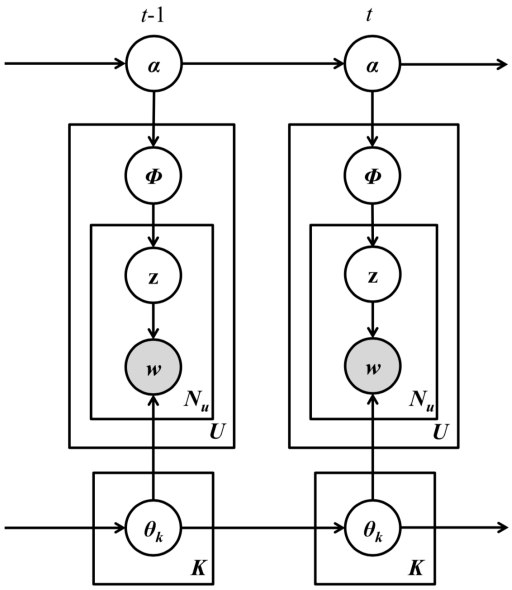

図 3　DTM のグラフィカルモデル

Teh らは，LDA の拡張として混合分布の事前分布に次数が可変である階層ディリクレ過程を用いることで，トピック数をパラメータとして与えずにトピック分布を推定可能な Hierarchical Dirichlet Process [4] を提案した．

また，Dynamic Topic Model (DTM) [2] は，Blei らにより提案されたもっとも代表的な時系列トピックモデルである．図 3 に DTM のグラフィカルモデルを示す．DTM では，単位時間毎に入力データと潜在変数を導入している．混合分布の事前分布であるディリクレ分布のパラメータは，直前のパラメータを平均とした正規分布からサンプリングされる．LDA や DTM の拡張モデルも数多く提案されているが [3], [5], [6], [7], [8]，いずれも各トピックは独立に発展していくと仮定していること，一括学習であることなどから，新しいトピックの発生や既存のトピックの混合を差分的にとらえることは困難であると考えられる．

2.2 トピックモデルの評価手法

トピックモデルの評価に際し，perplexity と coherence が指標として広く用いられている．perplexity はモデルの汎化性能を測る指標であり，学習済みモデルにおける単語群の予測尤度を

正規化することで与えられる．多くのトピックモデルの評価に用いられている一方で，perplexity の優れたモデルであっても一概に解釈性が高くなるとは限らず，perplexity が人間の評価に即さない可能性が Chang らによって指摘されている．[9] このため Chang らは各トピックの解釈しやすさを測る評価指標として coherence を提案した．「人から見た解釈のしやすさ」という曖昧な定義のため，これ以降，計算効率や精度の向上を図り，数多くの coherence 計算手法が提案されている．[10] [11] [12] [13]

本稿で提案する進化的トピックモデルでは，新しいトピックの発生に対して，突然変異や既存のトピックの交叉により発生したものと仮定することで，混合分布を推定する．本手法の実現に向け，本稿では perplexity，coherence に基づく時系列テキストデータに対するトピックの変化点検知に焦点を当てた考察を行う．逐次的に受け取られる新しいテキストデータに対し，学習済みモデルを用いて計算される perplexity を観察することで，新しいトピック分布の必要性を判定し，coherence を目的変数とした既存トピックの交叉などによる新しいトピック分布の生成することを目指す．

3 提案手法

3.1 概　　要

本研究で実現を目指す進化的トピックモデルでは，新しいトピックの発生に対して，突然変異や既存のトピックの交叉により発生したものと仮定することで，トピック分布を推定する．本手法の実現に向け，本稿では時系列テキストデータに対するトピックの変化点検知に焦点を当てた考察を行う．逐次的に受け取られる新しいテキストデータに対し，学習済みモデルを用いて計算される perplexity，coherence を観察することで，新しいトピック分布の必要性を判定し，既存トピックの交叉と突然変異により新しいトピック分布の生成することを目指す．

本手法の概要を図 4 に示す．本研究では時系列テキストデータを入力に LDA によりトピック分布・単語分布の推論を行い，これらを用いたテストデータに対する最尤推定により得られたトピック分布から文書毎の perplexity を計算する．この perplexity の値の時間経過に伴う推移をもとに，モデル更新時期の検知を行う．その後，モデル学習後に生起した新しい文書に対し，突然変異ないし既存トピック分布の交叉等により対応するトピック分布を作成し，モデルを更新する．更新したモデルを用いて同様の手順を繰り返し続けることで，出現し続ける新しい文書やトピックに対応し進化し続けるトピックモデルを実現する．

3.2 モデル更新時期の検知

多くの新しいテキストデータに対し，学習済みモデルのあてはまりが悪く perplexity が高い時，モデルを更新するべきだと考えられる．本稿では新しいテキストデータに対し，文書毎に perplexity を計算し，perplexity が上昇トレンドにある時点を検知することでモデルの更新を試みる．

訓練データ $\mathbf{w}_{train}$ を用いて学習された言語モデルにおける，単語群 $\mathbf{w}_{test}$ に対する perplexity は以下のようにして求められる．

$$perplexity(\mathbf{w}_{test}) = \exp(-\frac{\sum_{w_i \in \mathbf{w}_{test}} \log p(w_i \mid \mathbf{w}_{train})}{|\mathbf{w}_{test}|}) \quad (1)$$

$p(w_i \mid w_{train})$ の計算方法は言語モデルにより異なる．LDA の場合，単語の生起確率はトピックに，トピックの生起確率は文書に依存するため，単語 w の生起確率 $p(w)$ は以下のようにして求められる．

$$p(w) = \sum_t p(w \mid t) p(t \mid d_w) \quad (2)$$

ただし，t はトピック，d_w は w を含む文書を指し，$p(w \mid z), p(z \mid d_w)$ はそれぞれ LDA により推論されるトピック-単語分布，文書-トピック分布である．この perplexity の値が小さいほど，テストデータに対する言語モデルのあてはまりがよいとされる．

代表的な時系列トピックモデルにおいても，単語分布は短い期間の間に大きく変化することは少ないことを仮定している [2], [3]．したがって，新しい文書に対して学習済みモデルにおける単語分布を用いて perplexity を計算することは妥当であると考えられる．

一方で，LDA においては文書毎にトピック分布の学習を行うため，未知の文書に対するトピック分布を得ることはできない．したがって本研究では，学習済みモデルにおける単語分布を用いて新しい文書の尤もらしいトピック分布を推定する．このトピック分布を用いて文書毎の perplexity を計算し，この経時変化を観察することで，新しい文書に対するあてはまりが悪くなり，トピック分布の更新が必要とされるような時点の検知を試みる．

3.3 トピック分布の更新

最尤推定により得られたトピック分布を用いて計算された文書毎の perplexity の値が大きくなる場合，新しい文書に対するモデルのあてはまりが悪く，更新するべきであると言える．本研究では図 1 のような全く新しい話題については突然変異，既知の話題の新しい組み合わせについては交叉により，対応するトピック分布の生成を試みる．

最尤推定により得られたトピック分布を用いて計算された perplexity の値が大きくなる場合，再推定された単語分布に変化が少なければ，既存トピックとの干渉の少ない全く新しいトピックが，あるいは，既存トピックに類する特徴のみを持ち，新しい特徴を全く持たないものの，完全に一致もしないトピックが発生している可能性がある．この場合，前者を突然変異，後者を交叉により生成し，モデルの更新を行う．perplexity の値の変化は，単純な値の大小比較の他，確率過程によるモデリング等で捉えられると考えられる．

次に，推定された単語分布と学習済み LDA における単語分布が著しく異なる場合，トピック内での単語の使用数が変化していると考えられる．この場合，Dynamic Topic Model による推論，あるいは既存トピックの交叉による新トピックの作成により，モデルの経時変化を捉えた学習が可能になると考えられる．単語分布間の差異は，KL-divergence 等により捉えられる．

また，これらの両方に当てはまる場合，全く新しいトピック

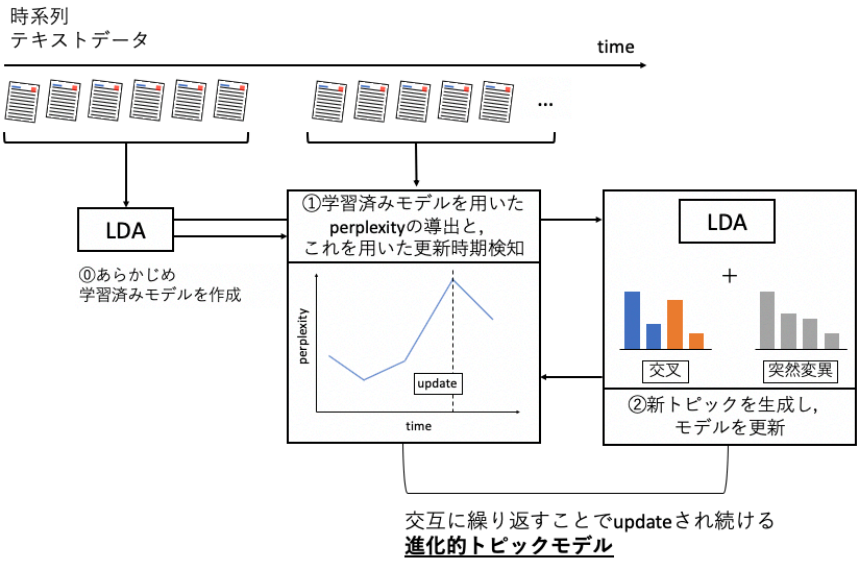

図 4　進化的トピックモデルの概要

が発生し，既存トピックとの干渉の起こった結果，学習済みモデルでは新トピックをとらえられていないため perplexity の値が大きくなり，単語分布も変化したと考えられる．この場合，新トピックを突然変異により生成し，モデルの更新を行う．

なお，本稿ではモデル更新時期の検知手法に焦点を当てた検証を行うため，新しいトピック分布の作成には上述の手法ではなく，便宜的に学習データとトピック数を更新した LDA を用いる．

4　実　　験

4.1　準　　備

本研究では実際のニュース記事を用いて進化的トピックモデルの実現に向けた実験を行なった．本実験では，読売新聞社における 2017 年 1 月から 12 月までのニュース記事 320531 件を用いた．古い順に 301045 件を訓練データとして LDA による推論に使用し，残る 19486 件のニュース記事について，トピック分布の推定と perplexity の計算を行なった後，単語分布の再推定を行なった．前処理として，全てのニュース記事について Mecab [14] を利用した形態素解析により普通名詞を抽出し，ストップワードを取り除いた．LDA のトピック数は DTM に倣い 20 とし，ハイパーパラメータ設定は $\alpha = 0.01$, $\beta = 0.01$，キギブズサンプリング行回数は 200 回とした．

4.2　トピック分布の推定と **perplexity**・**coherence** の算出

19486 件のニュース記事について，LDA により推論された単語分布 $p(w \mid z)$ を用いてトピック分布の推定を行う．単語群 $\mathbf{W}$ を持つニュース記事 d に対し，LDA の同様の生成仮定をおくと，ニュース記事の尤度 $p(W)$ は以下のようにして与えられる．

$$p(\mathbf{W}) = \prod_{w \in \mathbf{W}} \sum_{z} p(w \mid z) p(z \mid d) \qquad (3)$$

本稿では Stan 言語で，Hamiltonian Monte Carlo Sampling によって尤度 p を最大化する $p(z \mid d_w)$ の推定を行った．

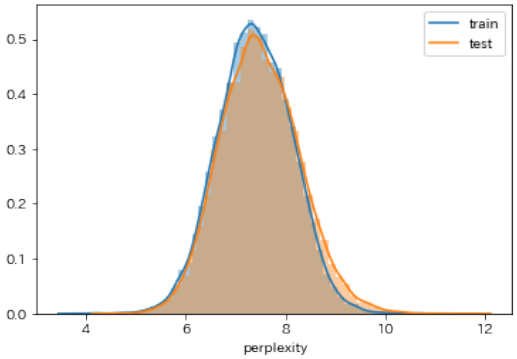

図 5　訓練・テストデータにおける perplexity の分布

図 5 は，推定されたテストデータの文書-トピック分布と LDA の推論結果を用いて式 1 から計算された，訓練・テストデータにおける文書毎の perplexity の分布を示す．LDA により計算された perplexity とほぼ同様の分布を持つことから，LDA と比べても大差ない文書-トピック分布を推定できていると言える．

また，図 6，7 は推定された文書のトピック分布の test-perplexity，移動平均 m の推移とその上昇・下降トレンドの期間を示す．いずれも縦軸は perplexity を，横軸はテキスト毎に古い順に割り振られた ID を示す．移動平均の区間幅は文書数の日平均に合わせ 900 とした．本実験では半区間で 0.08 以上 perplexity が上昇した場合上昇トレンドと，それ以外で下降トレンドと定義した．

図 7 と 8 を比較すると，perplexity に対応した変化を示すトピック分布があることがわかる．図 8 において際立った相関を示す 3 つのトピックを図 9 に示す．

これらトピック 3,6,12 について，単語分布 $p(w \mid t)$ や relevance($\lambda p(w|t) + (1-\lambda)p(w \mid t)/p(w)$，ただし $p(w)$ は入力データ全体での単語頻度) [15] の高い 5 個の単語を表 1 に，各

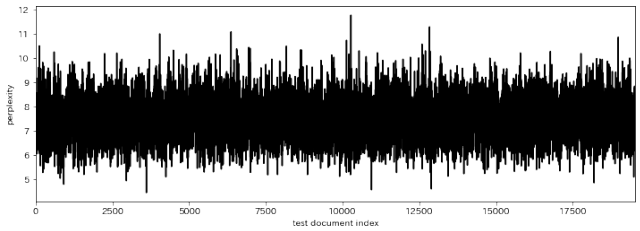

図 6　文書毎の perplexity の推移

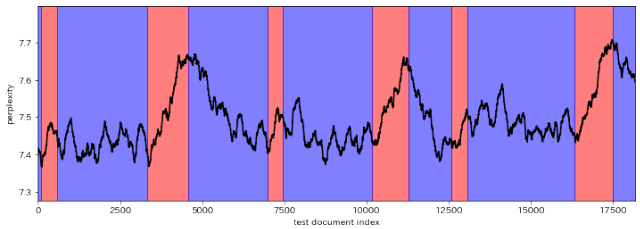

図 7　文書毎の perplexity の移動平均線

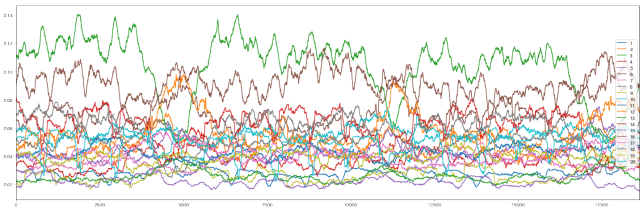

図 8　文書毎のトピック分布の推移

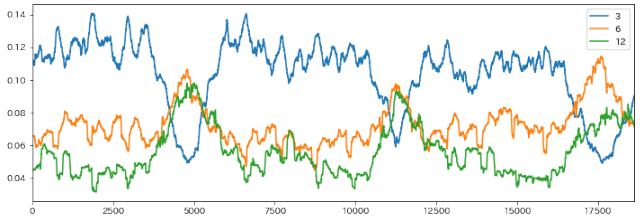

図 9　トピック 3,6,12 の文書毎の分布推移

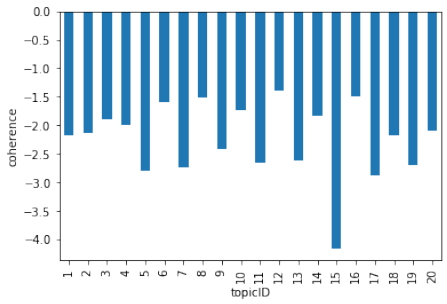

図 10　各トピックの coherence

トピックの UMass coherence [10] の値を図 10 に示す．これらによると，当該トピックはいずれも素性のわかりやすいトピックであり，coherence も相対的に高い値を示している．

このようなトピック-単語分布の推移が perplexity の推移に追従していることを踏まえると，coherence の高い高品質なトピックについて，これらの生起確率の高い文書の増減が perplexity の推移に反映されていると考えられる．

一方で，これら以外のトピックにおいては，coherence の値が悪い，解釈性の低いトピックも多い．そこで新トピックを生成することで coherence の改善が期待される．また，perplexity が一時的に下がった後戻るということは，当該以外のトピックは当てはまりがあまりよくないとも考えられる．新トピックの生成によりトピックの品質を向上させることで，当てはまりのよいトピックが増え，全体の perplexity の改善も期待することができる．

4.3　新しいトピック分布の作成

本研究で実現を目指す進化的トピックモデルにおいては，常に増え続ける新しい文書，新しい話題，変化する話題に対する適切なモデリングを目標としている．このため，新たな混合分布の作成には，差分的な学習手法が必要とされる．具体的には 3.3 節で述べた遺伝的アルゴリズムを用いた手法などが検討されうるが，本稿では，モデル更新時期の検知手法に焦点を当てた検証を行うため，便宜的に学習データとトピック数を更新した LDA を適用することで，モデルの更新を行った．また，それぞれの perplexity，coherence を計測し，妥当性を検証した．

図 7 に示されるトレンドの各変化点について，変化点までの新しい文書を入力に加え，合計文書数が変化しないように古い文書を入力から除いた学習データに対して LDA による推論を行い，新たなモデルに対して perplexity，coherence による評価を行った．トピック数以外の LDA のパラメータ等は 4.1 節に準拠した．

トピック数に関して，perplexity が上昇トレンドにある場合，既存のモデルのトピックでは対応できていないことを意味しており，トピック数を増やすことで新たなトピックを作成し割り当てる必要がある．一方で perplexity が下降トレンドにある場合，新しい文書に対する既存のトピックのあてはまりが改善されていることから，ある既の分布では対応できないトピックが用いられなくなったと考えられる．使用されるトピックが減っていることから，トピック数を減らすべきだと考えられる．

以上より本節では，下降トレンドの際はトピック数を減らし，上昇トレンドの際はトピック数を増やした LDA によりモデルを更新する手法について検証を行う．トピック数の逐次変更や，トピック数を減らす効果について確認するため，本手法に対するベースラインとして以下による推論も併せて行う．

（1）　トピック数を常に 20 とした LDA
（2）　トピック数を常に 19 とした LDA
（3）　トピック数が 20 から，上昇トレンド毎に 1 増える LDA

図 11，12 は各トレンド変化点までの文書を元に学習された LDA の perplexity，coherence を，表 2 はその平均値を示す．これらについて特にベースライン（3）に注目すると，トピック数を増やすと，perplexity は下がり，当てはまりが良くなっていると言える．この理由としては，モデルの持つ分布の個数が増えることによる表現能力の向上が考えられる．一方で，トピック数を増やすと coherence は全体として下がり，トピックの品質が悪くなる傾向が見られる．これはトピック数を増やすために必要以上にトピックが分割されているためと考えられる．

表 1　トピック 3,6,12 の単語分布 $p(w \mid t)$，relevance の上位 5 単語

	$p(w\|t)$	relevance($\lambda = 0.25$)	relevance($\lambda = 0$)
トピック 3	県，会，者，市，委員	県，市，職員，知事，会	未踏，継投，部局，石巻，奨学
トピック 6	の，人，こと，私，写真	私，映画，ん，言葉，父	ドラマ，劇，映画，演劇，主人公
トピック 12	選手，大会，チーム，戦，野球	選手，大会，チーム，野球，戦	選手，野球，リーグ，投手，サッカー

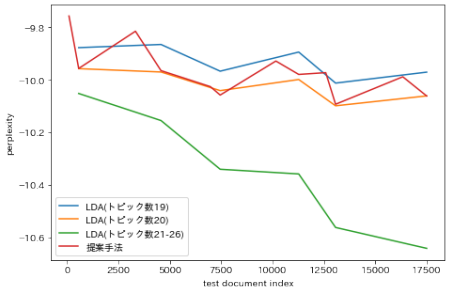

図 11　新しい文書を用いて再学習されたモデルの perplexity の推移

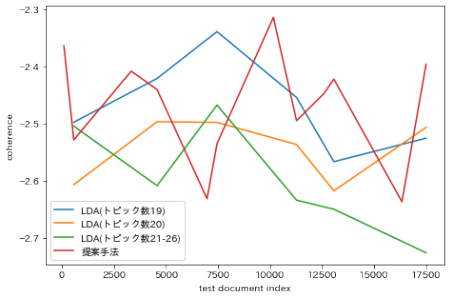

図 12　新しい文書を用いて再学習されたモデルの coherence の推移

表 2　再学習モデルの perplexity，coherence 平均値

	perplexity	coherence
LDA(トピック数 19)	-9.93	**-2.47**
LDA(トピック数 20)	-10.02	-2.54
LDA(トピック数 21-26)	**-10.35**	-2.60
提案手法	-9.97	**-2.47**

しかし，提案手法は perplexity，coherence ともに LDA(トピック数 19) 以上の値を示している．トピックの品質をなるべく下げずにトピック数を増やすことができており，モデルの更新に際しては，トピック数の固定ないし単調増加に比べ，提案手法のような時間毎のトピック数の調整は，高水準な perplexity，coherence を両立しうる有効な手法であると言える．

一方で，本実験では単純な方法でトレンドを決定したため，今後は TBSM [16] や短期・長期移動平均線を用いたトレンド分割についても検討したい．また各トレンドに対するより柔軟なトピック数の変更方法についても考察を深めていきたい．

5　ま　と　め

本稿では時間経過に伴う新しいトピックの発生や既存のトピックの混合を推定する進化的トピックモデルを提案した．提案手法は，文書毎の perplexity の時間経過に伴う推移をもとにしたモデル更新時期の検知と，新しいトピック分布を作成することによるモデルの更新を繰り返し続けることで，出現し続ける新しい文書やトピックに対応し進化し続けることを目指す．

perplexity，coherence に基づくモデル更新時期検知手法に焦点を当て，LDA によるモデルの更新を行った実験では，トピック数を単調増加させると coherence が悪化すること，適切なトレンド分割とトピック数の調整によって高い coherence を持つモデルが得られうることがわかった．

今後は実験を重ねモデル更新時期検知手法を確立するとともに，遺伝的アルゴリズムを導入した分布生成手法についても考察を行い，進化的トピックモデルの実現に向けた検証を進めていく予定である．

6　謝　　辞

本研究の一部は総務省 SCOPE(172307001) による．

文　　献

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *JMLR*, Vol. 3, pp. 993–1022, 2003.

[2] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd ICML*, pp. 113–120. ACM, 2006.

[3] 岩田具治, 渡部晋治, 山田武士, 上田修功. 購買行動解析のためのトピック追跡モデル. 電子情報通信学会論文誌 D, Vol. 93, No. 6, pp. 978–987, 2010.

[4] Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in NIPS*, pp. 1385–1392, 2005.

[5] Amr Ahmed and Eric Xing. Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering. In *Proceedings of the Conference on SIAM*, pp. 219–230, 2008.

[6] Amr Ahmed and Eric P Xing. Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. *arXiv preprint arXiv:1203.3463*, 2012.

[7] 佐々木謙太朗, 吉川大弘, 古橋武ほか. 複数のトピックの時間的依存関係を考慮した時系列トピックモデル. 研究報告数理モデル化と問題解決 (MPS), Vol. 2014, No. 3, pp. 1–6, 2014.

[8] Kentaro Sasaki, Tomohiro Yoshikawa, and Takeshi Furuhashi. Online topic model for twitter considering dynamics of user interests and topic trends. In *Proceedings of the Conference on EMNLP*, pp. 1977–1985, 2014.

[9] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in NIPS*, pp. 288–296, 2009.

[10] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the conference on EMNLP*, pp. 262–272, 2011.

[11] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd WWW*, pp. 1445–1456, 2013.

[12] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic

model quality. In *Proceedings of the 14th EACL*, pp. 530–539, 2014.

[13] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the 8th ACM*, pp. 399–408, 2015.

[14] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 conference on EMNLP*, 2004.

[15] Carson Sievert and Kenneth Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pp. 63–70, 2014.

[16] Jheng-Long Wu and Pei-Chann Chang. A trend-based segmentation method and the support vector regression for financial time series forecasting. *Mathematical Problems in Engineering*, Vol. 2012, , 2012.