文書内のトピック数を考慮したトピック追跡の試み

芹澤 翠　　小林 一郎
お茶の水女子大学人間文化創成科学研究科理学専攻
{serizawa.midori, koba}@is.ocha.ac.jp

1　はじめに

本研究では，新聞記事に存在するトピックを追跡し，時間変化に伴う事象の変化を分析することを目的とする．一文書に複数のトピックが含まれることは多くあり，これは新聞記事に関しても例外ではない．そのため，トピック追跡についても文書という単位でトピックを捉えるのではなく，対象文書集合全体に存在するトピックを対象とする必要があると考えられる．多くのトピック追跡についての研究では，トピックを文書クラスタリングを用いて抽出しており，上述のような前提をもっていない．今回はトピックの抽出に一文書に複数のトピックが存在することを表現できる Latent Dirichlet allocation(LDA) を用いる．LDA を用いる際には，文書内のトピック数を与える必要があるが，それを事前に決めることは困難である．その問題を解決するための方法としては，パープレキシティを利用しモデルを評価することでトピック数を決定する方法や HDP-LDA[12] を利用する方法が考えられる．一方，本研究では，トピックの内容の類似度に着目してトピック数を決定した上でトピックを抽出する手法を提案し，トピックの追跡を試みる．

2　関連研究

時系列のテキストデータを対象にしたトピック抽出およびトピックの発展を追跡するための手法は様々に提案されている [1, 2, 3, 4, 5]．テキストデータを対象にしたトピック抽出方法としては，文書クラスタリングを行い，抽出された文書のクラスタをトピックと見なす手法が多く用いられている．具体的には，階層型クラスタリングにおいて，語の共起性を考慮した方法 [1]，単語によって特徴付けられた文書ベクトルの類似度を利用する方法 [2, 3, 4, 5] などがある．トピック追跡においては，トピックの時系列連鎖に着目した手法として，トピック抽出のための文書クラスタリング内で日時を考慮するもの [3, 4]，隣接する期間ごとのトピックの類似度に基づき関連付ける方法 [5]，時制クラスタ内のトピック類似度を基づき関連付ける方法 [1] などがある．これらの研究と本研究との相違点として，いずれの研究においても，追跡対象となるトピックの単位を文書集合として捉え文書中にトピックが細分化されているという前提を持たないことが挙げられる．時系列性を考慮した LDA の研究としては，トピックの時間発展を多重スケールで捉えるモデルの提案などがされている [6]．この提案モデルでは，一時刻前の多重スケールパラメータをモデルに組み込むことで，トピックの時間発展が考慮されるように工夫している．一方で，時系列に沿った潜在トピック数の変遷に関しては触れられていない．他にもタイムスライス間のトピックを類似度に基づいて繋げることで抽出したトピックチェーンという概念を用いた研究 [9] もある．本研究では，時間変化に沿った対象文書内のトピック数の変化を考慮したトピック追跡を試みる．

3　トピック抽出

3.1　文書の前処理

本研究では，トピックの抽出に LDA を用いる．LDA の処理対象には，形態素解析器 MeCab[1] により抽出した名詞を複合化処理した複合名詞と複合化処理されなかった名詞を処理対象とした．ただし，複合名詞は新聞社や記者によって同じ意味の語でも表現方法が異なる可能性があるため，本稿では，複合名詞の統一を対象期間内の全対象文書に対して，以下の規則に基づいて行った．

- サ変接続の名詞を含む場合は複合化処理を行わない
  例えば,「映像流出」と「映像が流出した」という表現はともに「映像」「流出」と表現される．
- 構成する名詞に表記上の包含関係がある複合名詞は，構成する名詞の語数の少ない複合名詞へ置き換える

例えば,「来年度政府予算案」と「来年度予算案」はいずれも「来年度予算案」と表現される．

3.2　Latent Dirichlet allocation

LDA[7] は，一文書に複数トピックが含まれることを表現できる，文書生成過程の確率的なモデルである．具体的には，次のような生成過程となる．まず，トピック集合の各トピックについてディリクレ分布に従い語彙の多項分布を選ぶ．次に，各文書についてディリクレ分布に従いトピック上に定義された多項分布を選ぶ．

[1] http://mecab.sourceforge.net/

最後に，文書中の各単語に対してこの多項分布に従ってトピックを 1 つ選び，そのトピックに対応する初めに選んだ語彙の多項分布に従って語彙を 1 つ選ぶ．この処理を文書を構成する単語の数だけ繰り返し，文書を構成する語彙を選択する．これは，語彙を 1 つ選ぶごとにトピックを選び直していることに等しく，そのため，1 つの文書に複数トピックが含まれることをモデル化できる．文書生成モデルを評価する指標としては，広くパープレキシティが利用されている [7]．処理対象となる文書 D_{test} の総数を M とした場合，パープレキシティは以下の式 (1) で計算される．

$$perplexity(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M} N_d}\right\} \quad (1)$$

3.3 トピック内の語の特徴量

ある文書内の語の重要度の尺度として，tf-idf 値が頻繁に用いられている．本稿では，tf-idf 値での文書をトピックに置き換えた term-score[8] を語のトピック内での特徴量として用いる．トピック k での語 v の term-score は，あるトピックの語の出現確率を $\hat{\beta}$ として，以下のように計算される．

$$\text{term-score}_{k,v} = \hat{\beta}_{k,v}\log\left(\frac{\hat{\beta}_{k,v}}{(\prod_{j=1}^{K}\hat{\beta}_{j,v})^{\frac{1}{K}}}\right) \quad (2)$$

$\hat{\beta}_{k,v}$ ： トピック k での語 v の出現確率
K ： トピックの総数

この式では，単語のトピック内の出現確率である $\hat{\beta}_{k,v}$ が tf 値に対応し，残りの部分は，全トピックで頻繁に現れる語には値が低くなるため idf 値に対応している．

3.4 トピックの類似度

抽出された各トピックを，そのトピック内の特徴語とその特徴量を各次元に対応付けたベクトルである，トピックベクトルで表現する．そして，トピック間の類似度をトピックベクトルのコサイン類似度によって測る．ベクトル $\vec{x_1}, \vec{x_2}$ のコサイン類似度とは，以下の式で計算される，2 つのベクトルのなす角度のコサイン値であり，値が大きいほど 2 つのベクトルの類似度が大きいと判断できる．

$$cos(\vec{x_1}, \vec{x_2}) = \frac{\vec{x_1}\cdot\vec{x_2}}{|\vec{x_1}||\vec{x_2}|} \quad (3)$$

3.5 トピック数の判定

LDA には，トピック数は既知であるという前提がある．一方，対象とするトピックは文書において陽に観測されない潜在的なものを扱う．LDA でのトピック数を決定する方法として，パープレキシティに基づく方法が考えられるが，パープレキシティは本質的にトピックモデルを用いて抽出されたトピックの内容を定量的に評価することは考慮できないという点がある [11]．他の方法としては，階層型ディリクレ過程 [12] を導入し，事前にトピック数を与えることなくトピックを推定する HDP-LDA を用いる方法が考えられる．

本稿では，抽出されたトピックに含まれる語の意味を重視してトピック数を決定する手法を採ることとした．意図的に大きめのトピック数でトピックを抽出し，抽出されたトピックを類似度により結合することで，対象とする文書に適したトピック数を決定する．

3.5.1 類似度に基づくトピック数の判定

大きめに設定されたトピック数の下で LDA を用いて抽出したトピックに対し各トピック間の類似度を求め，閾値[2]以上の類似度を持つトピック組を ‘類似トピック組’，その中に含まれていないトピックを ‘単独トピック’，類似トピックを 1 つのトピックとしてまとめて生成されるトピック集合を ‘結合トピック’，および複数の結合トピックに含まれるトピックを ‘重複トピック’ と呼ぶ．トピック数の決定手順を以下に説明する．

step 1. 単独トピックの判定
抽出された各トピックに対しトピックベクトルを付与し各ベクトル間の類似度を式 (3) により測り，類似トピック組と単独トピックを決定する．

step 2. 結合トピックの生成
各トピックをノードとし各類似トピック組の 2 トピック間にリンクを張ったグラフを考える．このグラフ中の完全グラフを構成するノードを 1 つにまとめたものを結合トピックとする．

step 3. 重複トピックの判定
生成した結合トピックを構成するトピックを 1 つずつ見て行き，2 つ以上の結合トピックに含まれるトピックを重複トピックと判定する．

step 4. トピック数の判定
重複トピックは主張性が低いと見なし，各結合トピックから削除する．この重複トピックを削除した後の結合トピックを ‘重複トピックを除いた結合トピック’ とする．そして，ここで得られた「単独トピック数」と「重複トピックを除いた結合トピック数」の和を文書の持つ潜在トピック数と判定する．

3.6 トピック数判定実験

実際のニュース記事を対象にトピック数判定実験を他手法と比較して行った．比較する手法は，既述した HDP-LDA を用いた方法およびパープレキシティに基づいた方法である．

3.6.1 実験仕様

対象とする文書はニュースサイト「YOMIURI ONLINE（読売新聞）[3]」,「毎日 jp（毎日新聞）[4]」から

[2] 閾値は，類似度の乖離に基づき決定される．
[3] http://www.yomiuri.co.jp/
[4] http://mainichi.jp/

表 1: 各手法により判定されたトピック数

日にち	手案手法	LDA	HDP-LDA
11 月 13 日	11	14	9
11 月 14 日	10	7	7
11 月 15 日	10	9	8
実行時間 (sec)	890.78	5852.84	926.15

キーワード「尖閣」を与えて収集した 2010 年 11 月 13 日から 15 日までの 3 日間の 86 件のニュース記事である．

モデルの推定方法はギブスサンプリングを用い，その反復回数は 200，結果は 5 回実験を行った平均を用いている．提案手法での実験におけるモデルのパラメータは $\alpha = 0.1, \beta = 0.1$ とし，初めに与えるトピック数は 18 とした．類似トピックを決定する類似度の閾値は，予備実験から 0.06 と設定した．パープレキシティに基づいてトピック数を定める方法では，パラメータは提案手法と同様にし，トピック数に 2 から 50 の値をそれぞれ用いて LDA によりトピックを抽出した．パープレキシティが低いほど適したモデルと見なせるため，パープレキシティスコアが低くなった際のトピック数を最適なトピック数と見なすこととした．HDP-LDA のパラメータは α は 0.1, γ は分布 $Gamma(1,1)$ に従うとした．また，基底分布のパラメータは 0.5 とした．

また，実験環境は以下の通りである．

- CPU ： Intel Core 2 Duo Processor (2.53 GHz)
- OS ： Microsoft Windows 7 Home edition 64 bit
- メモリ ： 4 Gbyte

3.6.2 結果

実験結果を表 1 に示す．トピック数については，手法間で大きな違いは見られない．また，実行時間の比較も，提案手法の結果は初期値に依存する事実を考慮する必要があるが HDP-LDA とほぼ同等であることが分かる．これらの結果から，トピックの内容を考慮した今回の提案手法は，トピック数と計算時間において，HDP-LDA に敵う結果をもたらすことが分かる．

4 トピック追跡

4.1 トピック追跡手続き

トピックの追跡は，連続する 2 日間の各トピックを類似度により関連付けを行う．3.1 節に述べた文書の前処理を対象期間の全対象文書に行った上で，以下の処理を行う．

step 1. トピック抽出

対象期間の各日において，以下の処理を行う．

1. トピック抽出 (1 回目)
   対象文書に対し，本来存在するトピック数より多めと思われる値をトピック数として指定し，LDA を用いてトピック抽出を行う．
2. トピック数の判定
   1. において抽出されたトピック集合に対し 3.5.1 節に詳述した方法により，対象文書の持つトピック数を判定する．
3. トピック抽出 (2 回目)
   決定したトピック数を指定し，再度 LDA によりトピック抽出を行う．ここで得られたトピック集合を対象文書の持つトピック集合とする．

step 2. トピック追跡

各日のトピック集合を対象に，連続する 2 日間の各トピック間の式 (3) で算出される類似度が閾値[2] 以上ならばトピック間に関連があるとすることで関連付けを行い，これを対象期間分繰り返す．

5 トピック追跡実験

5.1 実験仕様

使用するニュース記事，初めに与えるトピック数，および類似度の閾値には，3.6 節で用いたものを使用した．また，追跡のための類似トピック判定に用いる類似度の閾値は予備実験から 0.17 とした．判定されたトピック数は，5 回試行した結果の平均値を用いた．

5.2 結果

トピック抽出結果と追跡結果をそれぞれ表 2，図 1 に示す．なお，トピック抽出結果のトピックのラベルは，実験から得られた各文書のトピック混合分布を元に著者が付与した．対象期間は「尖閣諸島での漁船衝突映像の流出問題」や「APEC の開催」のあった時期であり，抽出されたトピックはこれらの話題を中心としていることが分かる．

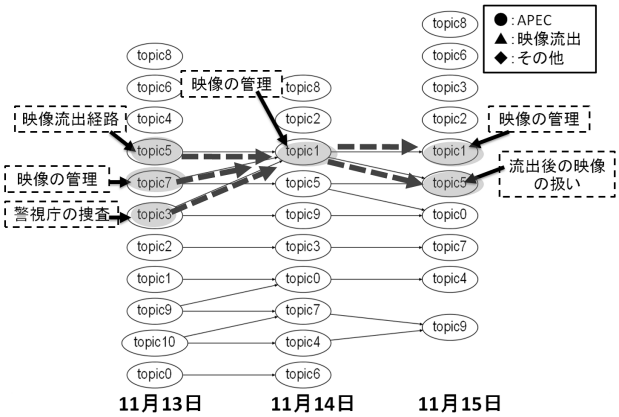

図 1: トピック追跡結果

5.3 考察

「映像流出」「APEC」とは関連の無いと思われる話題を「その他」とし，ラベルから各トピックをこの 3 つのカテゴリに分類して考察を行う．まず，表 2 から，13 日の topic8（民主党方針先送り）や 15 日の topic2 と topic7（福岡市長選）のように，その日のトピック

表 2: トピック抽出結果 (term-score 上位単語)

	トピック	term-score 上位単語	ラベル	カテゴリ
13日	topic0	映像 情報 国民 公開 政府	世論, 海保の見解	映像流出
	topic1	地域 中国 アジア 太平洋 発展	中国国家主席の講演	APEC
	topic2	馬淵 決議 責任 自民党 野党	大臣の責任問題	映像流出
	topic3	捜査 映像 検察 海保 逮捕	警視庁の捜査	映像流出
	topic4	人々 多大 私 弁護士 おかげ	航海士のコメント	映像流出
	topic5	映像 捜査 パソコン 神戸 警視庁	映像流出経路	映像流出
	topic6	前原 外相 提出 決議 不信任	不信任決議	映像流出
	topic7	映像 航海 パソコン 海保 流出	映像の管理	映像流出
	topic8	削減 政治 定数 改革 参院	民主党の方針	その他
	topic9	会談 首脳 中国 関係 会議	日中首脳会談	APEC
	topic10	米 日 首相 関係 合意 問題	日米首脳会談	APEC
14日	topic0	中国 関係 発展 開催 外務省	日中外相会談	APEC
	topic1	映像 パソコン 捜査 海保 保存	映像の管理	映像流出
	topic2	捜査 衝突 警告 船 現職 領海 海	事件についての海保の意見	映像流出
	topic3	参加 チベット 投開票 デモ 在日	政府への批判	APEC
	topic4	問題 中国 日本 米国 ロシア	日露，日米首脳会談	APEC
	topic5	航海 取り調べ 海保 同級生 富山	航海士の人物像	映像流出
	topic6	映像 尖閣諸島 現場 警備 政府	尖閣諸島沖の警備	映像流出
	topic7	会談 首脳 日 首相 官 会議	日中首脳会談	APEC
	topic8	アジア 存在 米国 パートナー 輸出	首脳会談まとめ	APEC
	topic9	処分 逮捕 起訴 懲戒 検察	航海士の懲罰	映像流出
15日	topic0	逮捕 航海 捜査 取り調べ 方針	航海士の懲罰	映像流出
	topic1	捜査 映像 神戸 海保 中国	映像の管理	映像流出
	topic2	選挙 事務所 広報 政策 見直し	福岡市長候補者の訴え	その他
	topic3	大使館 中国 郷送 金属 ライフル	中国への批判	映像流出
	topic4	中国 合意 外相 再開 前原	日中外相会談	APEC
	topic5	映像 航海 投稿 削除 私用	流出後の映像の扱い	映像流出
	topic6	沖 集合 時間 政府 日本	日本への批判	映像流出
	topic7	高島 福岡 吉田 自民 民主 政権	福岡市長選結果	その他
	topic8	公開 衆院 与党 自民党 審議	不信任案の可決	映像流出
	topic9	首脳 会談 会見 政府 領土	日本政府の会談への見解	APEC

で中心となっている「APEC」「映像流出」とは大きく内容が離れている「その他」の話題は単独のトピックとして正確に取り出せていることが分かる．また，他のカテゴリに属するトピックについても，それぞれ内容が重複することなくトピックが取り出せていることが結果からわかる．これは，提案手法により，適切なトピックの抽出が出来たと解釈できる．しかし，提案手法はトピック数の判定を閾値により決定した類似トピックを結合することにより行っているが，この手法で用いられる類似トピックは閾値に依るため，適切な閾値を設定しなけらばならないという問題がある．そのため，閾値の設定方法に関してはさらなる検討が必要であると言える．

また，トピックの追跡については，追跡結果から「映像流出問題」と「APEC」についてのトピックがそれぞれ追跡できていることが分かり，詳しく見てみると，「映像流出問題」のトピック追跡に関しても「流出した映像」に関する話題のみが追跡されているなど，関連する話題のみが追跡できていることが分かる．

6 おわりに

本稿では，潜在的トピックに基づくトピック追跡をするために，LDA を用いたトピック抽出を行った．また，対象文書内のトピック数が未知である問題を解決するために，文書が本来持つであろうトピック数より多めに抽出したトピックを類似度により結合することによりトピック数の判定を行った．提案手法が適切なトピック数を判定できているかを確認するため，ニュース記事を用い，パープレキシティによりモデルを評価する方法と HDP-LDA を利用する方法の 2 つの方法と決定されたトピック数についての実験と考察を行い，提案手法が HDP-LDA と同程度の結果をもたらすことを確認した．そして，連続する 2 日毎のトピック間類似度に基づいてトピックを関連付けることによりトピックの追跡を行い，実験により本提案手法の検証を行った．

今回はトピックの類似という観点からアプローチを行ったが，閾値により類似性を判定していたため結果が閾値によってしまうという問題があった．今後の課題としては，閾値によらずにトピック数を判定し追跡する方法の検討を考えている．

参考文献

[1] 森 正輝, 三浦 孝夫, 塩谷 勇, “時制クラスタのトピック追跡”, DEWS2006 論文集, 6A-i5, 2006.

[2] 平田 紀史, 児玉 政幸, 伊藤 正都, 大囿忠親, 新谷 虎松, “ニュース記事閲覧のための複数ウィンドウ方式を用いた特定トピック追跡システムの試作”, 全国大会講演論文集 第 70 回, ”1-633”-”1-634”, 2007.

[3] 菊池 匡晃, 岡本 昌之, 山崎 智弘, “階層型クラスタリングを用いた時系列テキスト集合からの話題推移抽出”, 日本データベース学会論文誌 Vol.7, No.1, pp.85-90, 2008.

[4] 平田 紀史, 大囿 忠親, 新谷 虎松, “ユーザの選好に基づくトピック分析システムの試作”, 第 22 回人工知能学会 全国大会, 3G1-01, 2008.

[5] 水落 大史, 井上 悦子, 吉廣 卓哉, 村川 猛彦, 中川 優, “新聞記事集合に対する時系列のトピック抽出”, DEIM フォーラム 2010 論文集, D6-3, 2010.

[6] 岩田 具治, 山田 武士, 櫻井 保志, 上田 修功, “オンライン学習可能な多重スケールでの時間発展を考慮したトピックモデル”, 情報論的学習理論テクニカルレポート 2009, 2009.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, 3:993-1022, 2003.

[8] D. M. Blei, and J. D. Lafferty, “TOPIC MODELS”, In A. Srivastava and M. Sahami, editors, Text Mining: Theory and Applications. Taylor and Francis, 2009.

[9] Dongwoo Kim and Alice Oh, “Topic Chains for Understanding a News Corpus”, The 12th International Conference on Intelligent Text Processing and Computational Linguistics, Japan, Feb. 2011.

[10] Chan Wang, Caixia Yuan, Xiaojie Wang, and Wenwei Xue, “Dirichlet Process Mixture Models based Topic Identification for Short Text Streams”, Proceedings of the 7th IEEE Conference on Natural Language Processing and Knowledge Engineering Tokushima,Japan, Nov.27-29,2011.

[11] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin, “Automatic Evaluation of Topic Coherence”, The 2010 Annual Conference of the North American Chapter of the ACL, pp.100–108,California, Jun., 2010.

[12] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, David M. Blei, “Hierarchical Dirichlet Processes”, Journal of the American Statistical Association, Vol.101, 2004.