

潜在的ディリクレ配分法に基づくトピック類似度を考慮したトピック追跡

芹澤 翠[†] 小林 一郎^{††}

[†] お茶の水女子大学理学部情報科学科 〒112-8610 東京都文京区大塚 2-1-1

^{††} お茶の水女子大学大学院人間文化創成科学研究科理学専攻 〒112-8610 東京都文京区大塚 2-1-1

E-mail: [†]{g0720526,koba}@is.ocha.ac.jp

あらまし ニュース記事などの時系列データを対象に潜在的な意味を考慮したトピック追跡を行う手法として、潜在的ディリクレ配分法(LDA)を用いたトピック追跡手法が提案されている。本研究では、対象となる文書における潜在トピック数を決定する際、LDAにて一度抽出したトピックに対して、トピック間の類似度を判定し統合することにより、適切なトピック数を求め、それに基づきトピックを抽出し追跡を行う手法を提案する。

キーワード 潜在的ディリクレ配分法, トピック追跡, 潜在トピック数

Topic Tracking based on Topic Similarity with Latent Dirichlet Allocation

Midori SERIZAWA[†] and Ichiro KOBAYASHI^{††}

[†] Department of Information Science, Faculty of Sciences, Ochanomizu University

2-1-1 Ohtsuka, Bunkyo-ku, Tokyo, 112-8610 Japan

^{††} Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

2-1-1 Ohtsuka, Bunkyo-ku, Tokyo, 112-8610 Japan

E-mail: [†]{g0720526,koba}@is.ocha.ac.jp

Abstract As a method to track topics considering the latent semantics of time-series text data, e.g., news articles, several topic tracking methods using Latent Dirichlet Allocation (LDA) have been proposed. In this study, we propose a method to estimate a proper number of topics in objective documents when the latent topics of the documents are analyzed by means of LDA — in concrete, in terms of deciding the number of latent topics in the objective documents, we firstly extract topics by means of LDA; unify the topics based on their similarity; and then regard the number of the unified topics as the relevant number of latent topics in the objective documents. By using this topic number, we reapply LDA to the documents and then track topics based on the extracted latent topics.

Key words Topic Tracking, Latent Dirichlet Allocation, Latent Topic Number

1. はじめに

我々の周りの事象は、時間の経過とともに内容が変遷していくことが多々ある。そのため、特定の時刻における情報は非常に断片的であり、断片的な情報のみでは、その事象の内容を包括的に把握することが困難である。このことから、ある事象に対し、時間的な内容の変化を捉えることで、全体像を掴み、その事象を深く解釈することが必要であると考えられる。そこで、本稿では、時系列に沿って取得された文書を対象に、ある特定のトピックの内容の変遷の分析を行う。特に対象文書として、ニュース記事を取り扱い、記事中の事象内容の変遷を分析する。

通常、一つの文書には複数の話題(トピック)が含まれている。ニュース記事もこの例外ではなく、単一の記事といっても

記載されている話題は1つでないことも多い。そのため、追跡の対象となるトピックには、文書という枠を超えた、対象文書全体に存在するトピックを対象とする必要があることが考えられる。これを実現するため、本稿では、従来トピック抽出に用いられていた文書クラスタリングではなく、結果として単語の持つ潜在的意味のクラスタリングとも解釈できる、確率的潜在意味解析によりトピック抽出を行い、潜在的トピックの追跡を行うことを目的とする。具体的には、潜在的ディリクレ配分法(LDA:Latent Dirichlet Allocation)を利用し、対象記事に存在する潜在的トピックを抽出する。一方、この方法を適用するにあたって、LDAの性質から、トピック数を予め指定する必要がある。しかし、文書内に潜在するトピックの数は未知であり、陽に決定することはできない。これを解決するため、潜在的ト

ピックの類似度から類似度の高いトピックの結合を行い、文書に潜む真のトピック数を決定する方法も提案する。

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]~[7]。トピック抽出においては、映像コーパス内の字幕テキストのような連続的なテキストデータのトピック分割方法としては、対象文を窓幅で区切り連続する窓幅間の類似度を評価し、最も評価の大きかった窓幅を採用することでトピック分割を行う方法がある [1]。また、ニュース記事のような離散的な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して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 [2]、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 [4]~[7] などがある。トピックの追跡においては、トピックの時系列連鎖に着目した手法としては、トピック抽出での文書クラスタリングにおいて日時を考慮するようにしたもの [5], [6]、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 [7]、時制クラスタ内のトピック類似度に基づき関連付ける方法 [2] などがある。また、単純なトピックの時系列連鎖に加え、トピック追跡結果を利用するユーザーの立場を考慮して時系列的に新しいトピックに古いトピックを依存させるようにスレッドを構築する手法 [1] も提案されている。そ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ており、文書中にトピックが細分化されているという前提を持っていないことが挙げられる。本研究では、文書中に複数のトピックが存在することを想定し、追跡対象を文書中の詳細なトピックとする。

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 [8]。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。本研究では、使用するトピックモデルに時間発展性は考慮できていないが、潜在的トピックの類似度に基づくトピック間の関連付けを行うことで、文書中のトピック数を判定し、トピックの時系列に沿った展開を捉える。

3. トピック抽出

3.1 対象文書の前処理

本研究では、トピックの抽出に LDA を用いる。LDA では、文書中に出現する一単語ずつを対象に処理を行うため、処理対象とする単語の種類は重要となる。他の品詞に比べ、意味を強く表す単語の種類として名詞が考えられるが、今回はより単語の意味を限定するために、形態素解析器 MeCab^(注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という問題がある。この問題を解決するため、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」という表現は意味上は同じだが、単純に複合化処理をすると、前者は「映像流出」となり、後者は「映像」「流出」と異なる表現となってしまう。一般的に、サ変接続の名詞は動詞「する」と接続して述語として使用される。そこで、サ変接続の名詞は複合化処理の対象から外すこととした。この例では「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
複合名詞に表記上の包含関係がある場合、通常、構成する名詞の語数が少ない複合名詞の方が文書中に出現する回数が多いと考えられる。例えば、「来年度政府予算案」と「来年度予算案」は同一の表現として処理されないが、後者の要素の名詞は前者に含まれているため、いずれも「来年度予算案」と表現される。

3.2 潜在的ディリクレ配分法

潜在的ディリクレ配分法 (LDA: Latent Dirichlet Allocation) [9] 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。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い。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。これは、トピックをクラスタと見なすと、単語単位のクラスタリングと見なすことができる。この文書生成過程では、潜在変数や未知パラメータを仮定しているが、実用的には、文書の単語分布から未知パラメータの推定を行う。本稿では、LDA の提案論文に則し、推定には変分ベイズ法を使用する。

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。tf-idf 値とは、語の一文書内での出現しやすさを表す tf 値と全文書中での語の希少度を表す idf 値の積を取り、その値が大きい方が語の文書内での重要度が高いとする尺度である。本稿では、この考え方に基づき、tf-idf 値での文書をトピックに置き換えた term-score [10] 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。

(注1): <http://mecab.sourceforge.net/>

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

$\hat{\beta}_{k,v}$: 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。

$$\cos(\vec{x}_1, \vec{x}_2) = \frac{\vec{x}_1 \cdot \vec{x}_2}{|\vec{x}_1| |\vec{x}_2|} \quad (2)$$

3.5 トピック数の判定

LDA の使用においては、トピック数は予め定められてい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 を単語単位のクラスタリングとみなし処理を行う際、トピック数の指定が必要となるため、その指定されるトピック数は重要である。そこで、本稿では、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 トピック抽出手続き

大きめに設定されたトピック数の下、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式 (2) により求め、閾値^(注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。

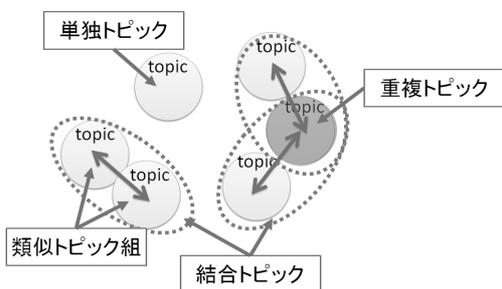


図1 トピックの名前

トピック抽出の処理手続きを説明する。LDA により抽出されたトピック集合について以下の処理を行い、単独トピック、重複トピック、および結合トピックを生成する。

step 1. 単独トピックの判定

LDA により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トピックベクトル間の類似度を式 (2) により測り、類似度が閾値^(注2)以上のトピック組を類似トピック組とする。決定した類似トピック組に一度も含まれないトピックを単独トピックとして決定する。

step 2. 結合トピックの生成

決定した類似トピック組について、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つのトピックとしてまとめたものを結合トピックとする。例えば、図2のようなグラフを考えると、点線円内に含まれるトピック集合が結合トピックに対応する。

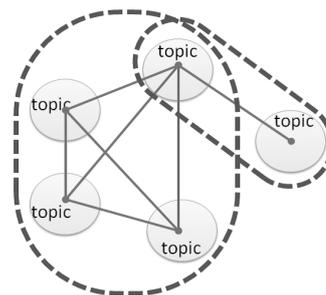


図2 結合トピックの例

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。

3.6 トピック抽出実験

トピック数の決定のため、実際に LDA によりトピック抽出を行い、トピックの性質を確認する実験を行った。

3.6.1 実験仕様

対象とするニュース記事は、ニュースサイト「YOMIURI ONLINE (読売新聞)^(注3)」、「毎日 jp (毎日新聞)^(注4)」からキーワード「尖閣」を与えて収集した2010年11月17日のニュース記事10件とした。LDA におけるモデル推定には変分ベイズ法を用い、抽出するトピック数は15とした。

3.6.2 実験結果

抽出したトピック、文書のトピック分布、および結合後のトピックをそれぞれ表1、表2、表3に示す。表2について、出現確率を有効数字3桁とした際に0.00となった値は空欄になっている。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。また、考察の便宜上、単独トピックには を、重複トピックには をトピック名に記し、重複トピックのラベルの後ろには重複しているラベルを持つトピックのトピック名を記した。

(注2): 閾値は、類似度の乖離に基づき決定される。

(注3): <http://www.yomiuri.co.jp/>

(注4): <http://mainichi.jp/>

表 1 抽出されたトピック (term-score 上位単語)

トピック	term-score 上位単語	ラベル
topic0	強化 午後 見直し 必要 海上保安庁	海上警察権見直し
topic1	実施 梶谷 木田 視聴 報告	尖閣ビデオ視聴調査隠ぺい
topic2	処分 検討 判断 放置 懲戒	尖閣映像流出での処分
topic3	問題 状態 管理 閲覧 処分	尖閣映像流出での処分 (topic2)
topic4	輸出 日本 話 中国 結果	中国の対日輸出
topic5	調査 流出 説明 報告 直後	尖閣ビデオ視聴調査隠ぺい (topic1), 尖閣映像流出経路 (topic7)
topic6	輸出 日本 企業 方針 中国	中国の対日輸出
topic7	結果 流出 調査 映像 前	尖閣映像流出経路
topic8	送信 受信 名目 双方 結局	尖閣映像拡散の原因
topic9	監視 搭載 東シナ海 強調 映像	中国の漁業監視船出港
topic10	送信 海保 担当 監視庁 場合	尖閣映像拡散の原因 (topic8), 尖閣映像流出経路 (topic7)
topic11	首脳 会議 日 会談 ゼロ	首脳会談
topic12	海保 ミス 担当 共有 範囲	尖閣映像拡散の原因 (topic8)
topic13	送信 海保 ネットワーク 受信 その間	尖閣映像流出経路 (topic7)
topic14	自民党 高島 候補 集会 街頭	福岡市長選挙

表 2 文書のトピック分布

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
doc0		0.01	0.12	0.11	0.02	0.03	0.02	0.03	0.17	0.01	0.16	0.01	0.13	0.17	
doc1	0.83				0.05		0.08	0.03		0.01					
doc2			0.40	0.21	0.01	0.02	0.02	0.04	0.01	0.02	0.04			0.20	0.02
doc3		0.01	0.01	0.07	0.02	0.03	0.02	0.03	0.25	0.01	0.21	0.01	0.10	0.24	
doc4	0.02	0.01			0.01		0.07	0.02				0.01			0.86
doc5	0.08									0.90		0.03			
doc6	0.02	0.04		0.05	0.10	0.31	0.07	0.35		0.04	0.01		0.01		
doc7	0.01			0.03	0.02		0.03			0.03		0.88			
doc8	0.07			0.02	0.47		0.42								
doc9	0.02	0.55		0.01	0.07	0.17	0.03	0.11		0.02	0.01				

表 3 結合後のトピック

単独トピック	(0) (9) (11) (14)
結合トピック	(1, 5)(10, 3, 12, 13)(8, 10, 12, 13)(2, 3, 12)(5, 7)(4, 6)
重複トピック	(3) (5) (10) (12) (13)

3.6.3 考 察

まず、文書のトピック分布を見ると、topic0 の doc1 でのみ高い出現確率を持っているなど、いずれも単独トピックはある特定の文書でのみ高い出現確率を持つことが分かる。どの単独トピックもある文書で 0.83 以上の出現確率を持っており、そのトピックが出現する他の文書に比べ非常に高い確率となっている。一方、重複トピックは、どの文書においても低い出現確率となっている。また、抽出されたトピックの具体的な内容を見ると、単独トピックはいずれも他のトピックとは異なる固有の内容を持っている。これに比べ、重複トピックは、topic3 が topic2 と同じ「尖閣映像流出での処分」を内容に持っているなど、全重複トピックが他の重複トピックでないトピックと同じ内容を持っていることが分かった。

単独トピックが、他のトピックと類似しておらず固有の内容を持っていることや特定の文書でのみ出現することを考えると、単独トピックは主張性の高い内容を持つトピックであると考えられる。一方、重複トピックはどの文書においても出現する頻度は低く、内容も他のトピックと重複していることから、単独トピックとは反対の傾向を持ち主張性の低い傾向を持つと考えられる。

3.7 トピック数の決定方法

具体的なトピック数の決定方法について説明する。まず、対

象文書に対して、初期トピック数に本来存在するであろうトピック数より大きめの値を指定して LDA によりトピック抽出を行う。次に、抽出したトピックから単独トピックと結合トピックを生成する。3.6 節の予備実験の結果から、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。もし重複トピックのみで構成される結合トピックがあった場合、この結合トピック自体が削除される。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。

最後に、ここで決定したトピック数を与え、再度、LDA によりトピックを抽出し、抽出したトピックをその日の文書の持つトピック集合とする。

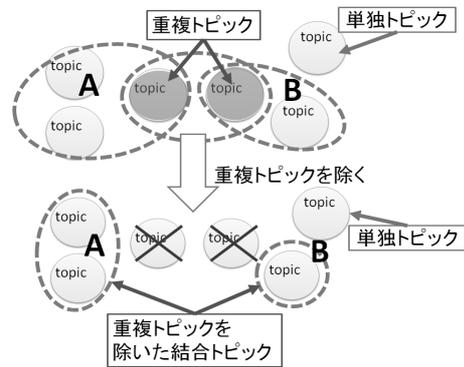


図 3 トピック数の決定

これにより、3.6 節の予備実験におけるトピック数の決定に用いる最終的なトピックは表 4 のようになり、この場合、決定トピック数は 9 となる。

表 4 トピック数の決定に用いる最終的なトピック

単独トピック	(0) (9) (11) (14)
重複トピックを除いた結合トピック	(1) (8) (2) (7) (4, 6)

4. トピック追跡

本研究においては、抽出した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 (2) で算出される類似度が閾値^{注2)}以上ならばトピック間に関連があるとすることで、関連付けを行う。この関連付けを対象期間の連続する 2 日ごとに行い、トピックを追跡する。また、前日から関連の付かなかったトピックは、その日に新たに生じたトピックと見なす。

4.1 トピック追跡手続き

本手法における、トピック追跡の流れを説明する。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。

表 5 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
11月13日	topic0	地域, 会談, 発展, 講演, 役割 中国国家主席の講演 日中首脳会談	APEC APEC
	topic1	パソコン, 保安, 捜査, 共用	映像の保存状況 映像流出
	topic2	私, 多大, 人々, 検察, おわび	保安官コメント 映像流出
	topic3	停泊, 胡, 接続, ネット, 保存	入手経路 映像流出
	topic4	主任, 航海, 事故, 海保, 管理	映像への海保の認識 映像流出
	topic5	米, 日, 合意, 米国, 首相	日米首脳会談 APEC
	topic6	責任, 馬淵, 決議, 野党, 問責	国交相の進退 映像流出
	topic7	削減, 政治, 定数, 改革, 歳費	民主党方針送り その他
11月14日	topic8	男性, 与党, 女性, 提出, 支援	不信任案検討 映像流出 国民の意見 映像流出
	topic0	再開, 交渉, 前原, 外相, 見方	日中外相会談 APEC
	topic1	民主, 参加, デモ, チベット, 米	国内政権への批判 APEC と 中国への批判 映像流出の両方
	topic2	現場, 警備, 調査, 領海, 警告	尖閣諸島の領海警備 映像流出
	topic3	会談, ギリギリ, 明確, 福山	日中首脳会談 APEC
	topic4	ロシア, 大統領, 米国, 日, 露	米露との首脳会議 APEC
	topic5	処分, 懲戒, 起訴, 逮捕, ケース	保安官の処分 映像流出
	topic6	主任, 航海, 同級生, 富山, 口数	保安官の人柄 映像流出
11月15日	topic7	パソコン, 保存, 共有, 同僚	映像の保存状況 映像流出
	topic0	削除, 私用, 自宅, 様子, 文書	流出後の映像 映像流出 保安官の様子 映像流出
	topic1	大使館, 郵送, 聴取, 海保, ライフル	映像の取り扱い 映像流出 中国への批判 映像流出
	topic2	意思, 取り調べ, 逮捕, 結論	保安官の処分 映像流出
	topic3	領土, 日, 時間, 集合, 会談, 国	首脳会談への政府見解 APEC
	topic4	提出, 決議, 不信任, 審議, 馬淵	不信任決議案提出 映像流出
	topic5	高島, 福岡, 吉田, 民主, 政権	福岡市長選 その他

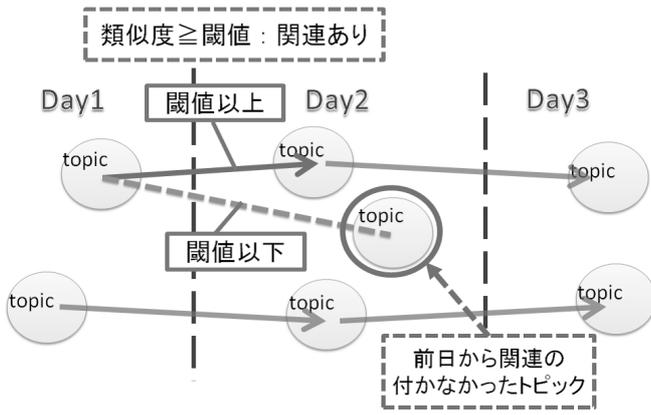


図 4 トピック追跡

1. トピック抽出 (1 回目)

対象文書に対し、本来存在するであろうトピック数より多めと思われる値を指定し、LDA を用いてトピック抽出を行う。

2. トピック数の決定

1. において抽出されたトピック集合に対し、各トピック間の類似度に基づいてトピックの結合を行う (トピックの結合については、3.5.1 節に詳述。)「重複トピックを除いた結合トピック数」と「単独トピック数」の和を、対象文書の持つトピック数とする。

3. トピック抽出 (2 回目)

決定したトピック数を指定して、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

抽出した各日のトピック集合について、各トピック間の類似度に基づき、連続する 2 日間のトピックの関連付けを対象期間分を行う。

5. 実験

数日間のニュース記事を対象に提案手法によるトピック追跡の実験を行い、本手法の妥当性を検証する。

5.1 実験の仕様

対象とするニュース記事は、ニュースサイト「YOMIURI ONLINE (読売新聞)^(注3)」、「毎日.jp (毎日新聞)^(注4)」からキーワード「尖閣」を与えて収集した 2010 年 11 月 13 日から 15 日までの 3 日間のニュース記事 86 件とした。LDA におけるモデル推定には、変分ベイズ法を用い、1 回目のトピック抽出で与えるトピック数は 15 とした。トピック数の決定における類似トピックを決める閾値は 0.060、追跡のための関連のあるとするトピック類似度の閾値は 0.144 に予備実験から経験的に設定した。また、決定トピック数は 1 回目のトピック抽出を 10 回繰り返した結果の平均とし、2 回目のトピック抽出結果は決定トピック数でのトピック抽出を 10 回行った内で抽出されたトピック集合内のトピック間類似度の最大値が最も小さい実

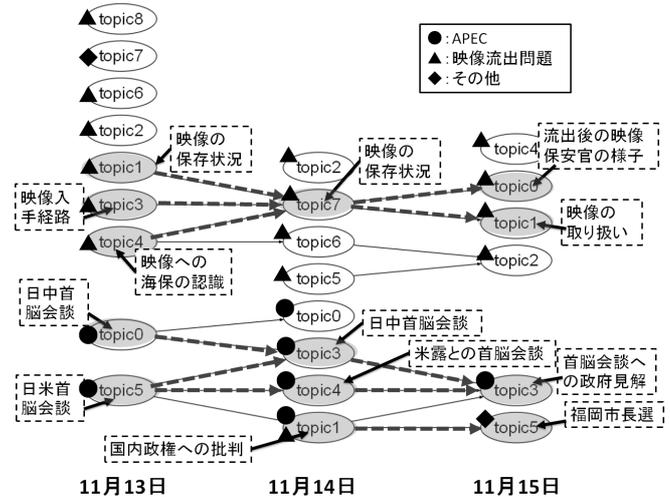


図 5 トピック追跡結果

行回のトピック集合を採用した。

5.2 実験結果

トピック抽出結果と追跡結果をそれぞれ表 5、図 5 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。

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。

a) トピック抽出について

11月13日の topic7 (民主党方針先送り) や 15日の topic5 (福岡市長選) のように、その日のトピック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。「映像流出」「APEC」に関する話題は、例えば13日の topic1 (映像流出:映像の保存状況) と topic2 (映像流出:保安官コメント) のように同じ「映像流出」に関するものでも概ね詳細にトピックが抽出できていることが分かる。一方、13日の topic0 のように同じカテゴリ「APEC」に属しているものの「講演」と「会談」についてと内容の異なる2つの話題が1トピックとして抽出されているものや14日の topic1 のように「映像流出」「APEC」の話題が混合されて「中国への批判」と「国内政権への批判」のトピックが1トピックとして抽出されているものもあった。前者は「中国国家主席」について、後者は「ある対象への批判」についての話題として抽出されたことが考えられる。この日の他のトピックの内容に重複が無いため、これらの話題を別のトピックとして抽出したい場合は、実際のトピック数はもう少し大きい値であると考えられる。トピック数の決定は、閾値により類似していると判定したトピックを結合することにより行っているため、閾値の設定に再検討を要することが考えられる。

b) トピック追跡について

まず、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おり、「APEC」のトピック追跡についても「会談」に関する話題のみが追跡できていることが分かる。他の話題に関しても、同じカテゴリに属する話題ごとに関連づけられていた。また、14日の topic2 は「尖閣諸島の領海警備」に関してのトピックであり、前後の日のいずれのトピックとも関連が付けられていない。トピック抽出結果から、13日のトピックにも15日のトピックにも映像流出についての話題はあっても「領海」や「警備」についての内容を持つものが無いため、この関連付けは妥当であり、このトピックはこの日に新たに生じたトピックと見なせると考えられる。一方、国政についての内容を持つ14日の topic1 (国内政権への批判) が地方都市の内容を持つ15日の topic5 (福岡市長選) に関連付けられているなど、一見関連の無いように思えるトピックが関連付けられているケースもあった。これら2つは直感的には関連の無さそうなトピックであるが、「政権」についての話題ということで関連が付いたのであると考えられる。このことより、潜在的なトピックの追跡をすることにより表層的には見えにくいトピックのつながりの発見をすることが可能であるとも考えられる。

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で、トピック数の決定を行った。具体的には、実験により、複数の結合されるトピックに含まれる重複トピックは、トピックとして主張性の低いものと見なすこととし、これらを除いた結合トピックと単独トピックのみを意味のあるトピックとして扱えるのではないかと考え、「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」の和をトピック数として決定した。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。それにより、ある程度の粒度ではトピック抽出はできているものの、詳細にはトピック抽出において再検討すべき余地があることが分かった。また、トピック追跡では、詳細な追跡に加え、表層的には見えにくいトピックのつながりの発見ができた。

今後の課題としては、今回の結果を踏まえたトピック数の決定方法の再検討、トピック追跡手法の改善、および抽出されたトピックのラベル付け方法の検討などが挙げられる。

文 献

- [1] 井手 一郎, 孟 洋, 片山 紀生, 佐藤 真一 “大規模ニュース映像コーパスの意味構造解析”, 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解 103(296), 13-18, 2003.
- [2] 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, 2006.
- [3] 張 一萌, 何 書勉, 小山 聡, 田島 敬史, 田中 克己 “時系列データに意味的に関連するニューストピックの発見”, DEWS2006 論文集, 2006.
- [4] 平田 紀史, 児玉 政幸, 伊藤 正都, 大園忠親, 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.
- [5] 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, 2008.
- [6] 平田 紀史, 大園 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, 2008.
- [7] 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, 2010.
- [8] 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.
- [9] D. M. Blei, A. Y. Ng, and M. I. Jordan “Latent Dirichlet Allocation”, Journal of Machine Learning Research, 3:993-1022, 2003.
- [10] 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.