潜在的なトピックの類似度に基づくトピック追跡

芹澤翠　(指導教員：小林一郎)

1　はじめに

ある事象を理解する際，時間的な内容の変化を把握することで，その全体像を掴み，深く理解することが可能になる．一方，単一の話題を取り上げていると思われる文書においても記載されている話題は複数のトピックにより構成されていることも多い．そのため，正確にその話題の経時変化を捉えるためには，その様な潜在的なトピックの追跡を行う必要がある．このことを考慮し，本研究では，確率的潜在意味解析によりトピックの抽出を行い，時系列データであるニュース記事を対象にトピック追跡を行うことを目的とする．

2　トピック抽出

本稿では，トピック抽出に潜在的ディリクレ配分法（LDA）[1] を用いる．これは，一文書に複数トピックが含まれることを表現できるトピックモデルであり，各文書は潜在トピックの混合分布として，トピックは語の確率分布として表現される．

2.1　トピック内の語の特徴量

LDA によって抽出したトピック内の語の特徴量として，以下の term-score が定義されている [2].

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

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

これは，tf-idf 値の考え方に基づいた尺度であり，単語のトピック内の出現確率 $\hat{\beta}_{k,v}$ が tf 値に相当し，語の 1 トピック内での出現しやすさを表しており，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 値に相当している．

2.2　トピック間類似度

本稿では，対象文書群の持つトピック数の決定および対象期間でのトピック追跡をトピックの類似度に基づいて行う．トピックの類似度は，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のコサイン類似度によって測る．

2.3　トピック数の決定

LDA は予め与えられたトピック数の下にトピックを抽出するが，トピック数は文書から陽に観測することはできない．そこで，本稿では，次のようにトピック数を決定する．まず，本来存在するであろうトピック数より大きめの値を意図的にトピック数として与えてトピックを抽出する．そして，抽出したトピックに対し，閾値[1]以上の類似度を持つトピック組を同じ内容を持つ ‘類似トピック’，その中に含まれていないトピックを ‘単独トピック’ と見なし，類似トピックを 1 つのトピックとしてまとめることで，‘結合トピック’ を生成する．ここで，本研究における先行研究 [3] により，複数の結合トピックに含まれるようなトピック（‘重複トピック’ と呼ぶ）を主張性の弱いトピックと捉え,「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」の和をその文書に潜在するトピック数と判定する．

[1]閾値は，類似度の乖離に基づき決定される．

3　トピック追跡

抽出したトピック群について連続する 2 日間の各トピック間の類似度が閾値[1]以上ならば関連があるとし，トピックを追跡する．前日から関連の付かなかったトピックは，新たに生じたトピックと見なす．以下に，本手法におけるトピック追跡の流れを説明する．

step 1. 対象文書の前処理
本稿では，名詞を複合処理した複合語と複合処理されなかった名詞を LDA によるトピック抽出の処理対象とした．複合語は新聞社や記者により同じ意味の語でも表現方法が異なる可能性があるため，複合語の統一を対象期間の全文書に対して次のルールに基づいて行う．

- サ変接続の名詞を含む場合は複合処理を行わない
- 構成する名詞に包含関係のある複合語は，構成する名詞数の少ない複合語へ置き換える

step 2. トピック抽出
対象期間の各日に対して，以下の処理を行う．

1. トピック抽出 (1 回目)
   文書に本来存在するトピック数より多めと想定されるトピック数を指定し，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群に対し，各トピック間の類似度に基づきトピックの結合を行う．結合後の「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」の和を対象文書の持つ潜在トピック数とする．
3. トピック抽出 (2 回目)
   決定したトピック数を指定し，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群を対象文書の持つトピック群とする．

step 3. トピック追跡
抽出したトピック群について，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行う．

4　実験

提案手法によるトピック追跡の実験を行い，本手法の妥当性を確認する．

4.1　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）」,「毎日 jp（毎日新聞）」からキーワード「尖閣」を与えて収集した 2010 年 11

表 1: トピック抽出結果 (term-score 上位語)

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

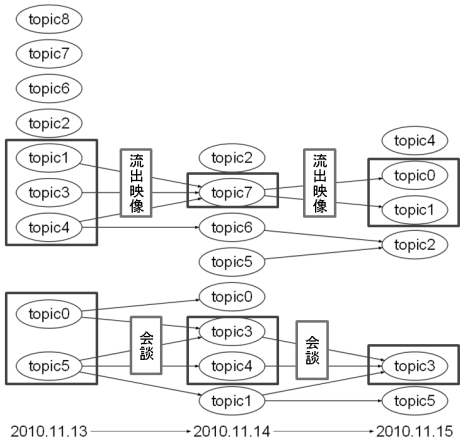

図 1: トピック追跡結果

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．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 回行った内でトピック間類似度の最大値が最も小さいトピック群を採用した．

4.2 トピック追跡結果

トピック抽出結果と追跡結果をそれぞれ表 1，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．

4.3 考察

「映像流出」「APEC」とは関連のない話題を「その他」とし，ラベルから各トピックをこの 3 つの話題に分類して考察を行う．以下，これらをカテゴリと呼ぶ．

まず，トピック抽出に関しては，11 月 13 日の topic7 や 15 日の topic5 のように,「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．一方,「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ものもあるが，13 日の topic0 のように同じカテゴリに属する 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 1 トピックとされているものもあった．これは，まとめられた 2 つの話題に共通して使用されている語が存在したために生じたものと考えられる．この日の他のトピックの内容に重複がないため，実際のトピック数はもう少し大きい可能性があり，決定したトピック数が最適ではなかったと判断できる．

トピック追跡に関しては,「流出映像」に関する話題と「会談」に関する話題は追跡ができている．他の話題に関しても，同じカテゴリに属する話題ごとが関連づけられていることが分かる．一方，14 日の topic1 が 15 日の topic5 に関連付けられているなど，さほど関連のな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．これは設定した閾値が適切ではないことから生じていると考えられ，トピック数の決定も閾値に基づき行っていることを考慮すると，閾値の決定方法の検討が必要である．

5 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実現するために，対象期間の各日において LDA を用いたトピック抽出を行い，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．また，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追跡を行った．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証とトピック追跡の改善，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．

参考文献

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

[2] 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.

[3] 芹澤翠, 小林一郎 “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”, DEIM Forum 2011, 2 月, 2011.