

潜在的な意味を考慮したトピック追跡の一考察

芹澤翠[†] 小林一郎[‡][†]お茶の水女子大学理学部情報科学科[‡]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情報科学コース

1 はじめに

ある事象の理解において、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、単一的话题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、正確にトピックを追跡するためには、その様な潜在的なトピックの抽出を行う必要がある。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]。

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

$\hat{\beta}_{k,v}$: 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。

A Study on Topic Tracking considering the Latent Semantics in a Document

[†]Midori SERIZAWA(g0720526@is.ocha.ac.jp),

[‡]Ichiro KOBAYASHI(koba@is.ocha.ac.jp)

[†]Dept. of Information Sciences, Faculty of Science, Ochanomizu University, 2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610

[‡]Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University, 2-1-1 Ohtsuka Bunkyo-ku Tokyo 112-8610

2.2 トピック数の決定

LDA ではトピック数は予め与えられているという前提があるが、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、トピック数に大きめの値を意図的に与えて抽出したトピックに対し、閾値*以上の類似度を持つトピック組を同じ内容を持つもの(‘関連トピック’と呼ぶ)とする。その中に含まれていないトピックを他のトピックと関連のない、‘単独トピック’とみなし、関連トピックを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。そして、単独トピック数と結合トピック数の和を結合後のトピック数とする。トピック間の類似度は、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(式(1)に表現される)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。

3 提案手法

本手法における、トピック抽出の流れを説明する。

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。ただし、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、複合語の統一を全対象文書に対して、次のルールに基づいて行った。

- 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 (1 回目)

対象文書群に対し、多めと思わ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。

step 3. トピック数の決定

step 2. 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いてトピックの結合を行う。結合後のトピック数を、対象文書群の持つトピック数とする。

step 4. 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群を対象文書群の持つトピック群とする。

*閾値は、類似度の乖離に基づき決定される。

4 実験

4.1 実験仕様

1日ごとのトピック抽出の実験を行い、本手法を用いて抽出したトピックの妥当性を確認する。使用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、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、トピック結合のために関連のあるトピック組は、類似度が閾値以上のトピック組を利用し、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た。

4.2 トピック数決定方法

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、最終的なトピック数を「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」との和とした。

4.3 結果

トピック抽出結果と結合後のトピック(いずれも繰り返し中1回分)、および2回目のトピック抽出結果をそれぞれ表1, 表2, 表3に示す。最終的な決定トピック数は9であった。また比較のため、恣意的にトピック数10として抽出したトピック(term-score 上位5単語)を表4に示す。なお、表1, 表3, 表4の各トピックのラベルは、著者が付与した。

2回目のトピック抽出結果では、topic0,7,8など尖閣映像流出にまつわるトピックが複数存在するが、詳細の内容が異なっている。一方、トピック数10として抽出したものはtopic0,1など内容に重複があるトピックも抽出された。

4.4 考察

2回目のトピック抽出結果から、抽出された9トピックはそれぞれ内容に重複のないトピックであると考えられる。一方、トピック数10として抽出された10トピックは内容に重複のあるトピックが含まれていた。このことから、対象記事群の持つトピック数は9であり、決定トピック数は妥当であったことが考えられる。

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するために、LDAを用いたトピック抽出を行い、文書群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、対象文書のトピック数の決定を行った。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、抽

表 1: トピック抽出 1 回目 (term-score 上位 5 単語)

トピック	term-score 上位 5 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べい
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
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べい, 尖閣映像流出経路
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 警視庁 場合	尖閣映像拡散の原因, 尖閣映像流出経路
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
topic13	送信 海保 ネットワーク 受信 その間	尖閣映像流出経路
topic14	自民党 高島 候補 集会 街頭	福岡市長選挙

表 2: 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)
重複トピックを除いた結合トピック	(1) (8) (2) (7) (4, 6)

表 3: トピック抽出 2 回目 (term-score 上位 5 単語)

トピック	term-score 上位 5 単語	ラベル
topic0	海保 担当 ミス 共有 連絡	尖閣映像拡散の原因
topic1	監視 搭載 東シナ海 受信 送信	中国の漁業監視船出港
topic2	強化 必要 見直し 午後 海上保安庁	海上警察権見直し
topic3	首脳 会議 会談 日 ゼロ	首脳会談
topic4	日本 輸出 話 調査 流出	中国の対日輸出
topic5	報告 視聴 木田 梶谷 実施	尖閣ビデオ視聴調査隠べい
topic6	自民党 高島 候補 支援 集会	福岡市長選挙
topic7	処分 判断 検討 可否 閲覧	尖閣映像流出での処分
topic8	送信 警視庁 受信 ネットワーク 担当	尖閣映像流出経路

表 4: トピック抽出 (トピック数 : 10)

トピック	term-score 上位 5 単語	ラベル
topic0	送信 警視庁 ネットワーク 担当 映像	尖閣映像流出経路
topic1	送信 受信 担当 依頼 ネットワーク	尖閣映像流出経路
topic2	報告 視聴 木田 実施 調査	尖閣ビデオ視聴調査隠べい
topic3	会議 首脳 会談 日 ゼロ	首脳会談
topic4	輸出 日本 話 方針 企業	中国の対日輸出
topic5	処分 閲覧 懲戒 問題 幹部	尖閣映像流出での処分
topic6	高島 自民党 候補 集会 支援	福岡市長選挙
topic7	処分 判断 検討 海保 対象	尖閣映像流出での処分
topic8	監視 東シナ海 搭載 漁業 午前	中国の漁業監視船出港
topic9	強化 見直し 午後 必要 海上保安庁	海上警察権見直し

出されたトピックのラベル付け方法の検討、およびトピック追跡の実装などが挙げられる。

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”, 第3回データ工学と情報マネジメントに関するフォーラム, 3月, 2011.