1P-2

文書内に潜在する事象の関係抽出に基づく俯瞰分析への取り組み

澤村瞳 † 小林一郎 †
† 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

近年，Web などにおける大量文書データの増加に伴い，膨大の情報の中から迅速に有益な情報を抽出する手法が必要とされている．それと共に，文書要約手法や文書内の情報を可視化するなどの手法が活発に研究されている．また，トピック分析など文書の表層的な情報だけではなく潜在的な情報を捉える手法も開発されてきており，様々な用途に用いられている．

本研究では，

文書全体の潜在的意味を解析し，トピックの変遷に伴う事象の生起から着目すべき情報や特定人物の行動履歴などを抽出する．これにより複数文書内に潜在する事象の関係抽出に基づく俯瞰分析手法を提案する．

2 関連研究

ニュース記事などを対象にした複数文書からトピック推定を用いて重要なイベントを抽出し，イベントの生起等を俯瞰する研究は広く行われている．Weiwei ら [2] は，ニュース記事から潜在ディリクレ配分法 (LDA) を用いてトピック推定し，類似しているトピックを抽出している．抽出したトピックをひとつに結合し，トピック毎のキーワードを追跡することで，イベントの動向を分析している．また，WenWen ら [3] は，トピック推定からイベントを抽出するのとは別に，人物や，場所となどの情報を抽出し，抽出したイベントと関連させることで，観測された情報をイベントと他の重要情報との関連性の観点から分析を行っている．

これらに対し，本研究ではイベント抽出においてトピック推定を行った後トピック毎に抽出された重要単語において，トピック間にわたり共起している単語の関係を調べることによりトピック間の相関関係を捉えることに着目する．

3 提案手法

3.1 概要

対象複数文書に対して潜在意味解析を行いトピックを抽出する．文書内の各文に対して推定されたトピックの中で一番重みの多いトピックをその文とトピックとみなす．文をトピックごとに分類することにより各トピックに対して具体的な表層情報を得る．それによりトピックを構成する重要単語の出現頻度を求め，各トピックおよび関連する情報の動向を捉える．

3.2 HDP-LDA

本研究では，文書の潜在的意味を階層ディレクレ過程を用いた配分法 (HDP-LDA: Hierarchical Dirichlet Process Latent Dirichlet Allocation) [1] により推定する．HDP-LDA は予めトピックを数を与えなくとも，自動でトピックを抽出する言語モデルであるため，与えるトピック数が推定結果に大きな影響を及ぼすことはない．

3.3 キーワード抽出

トピック推定した文集合から，特定の時期におけるキーワードを抽出する．Weiwei ら [2] によって提案された重要単語抽出手法（式 (1)，式 (2) 参照）を用いて，重みが大きい単語をキーワードとして抽出することで，その時に重要なイベントを認識する．式 (1) は前の時間を考慮し，他のトピックの単語出現度を考慮した式である．式 (2) はあるトピックのみに着目して，重要単語を抽出した式である．t は時間，k はトピック，TF は文中の単語出現度，λ は重みパラメータ，ISF は総文数/単語が出現する文を表す．

$$Weight(w)_k^t = \frac{TF(w)_k^t}{\sum_k TF_k^t} \times \exp(-\lambda \times Weight(w)_k^{t-1}) \tag{1}$$

$$Weight(w)_k^t = TF(w)_k^t \times ISF \tag{2}$$

3.4 人物の導入

人物や，国家，組織などをイベントに関連する情報とみなし，出現頻度からトピックやイベント等と，どのように関与しているかを抽出する．

4 実験

4.1 実験仕様

対象データは，毎日新聞の「911 テロ」に関する記事の 2001 年 9 月 12 日から 10 月 10 日までの 1438 件のニュース記事とする．

An Approach to Bird's-eye View Analysis based on extracting Relations among Latent Events in Documents

† Hitomi SAWAMURA(sawamura.hitomi@is.ocha.ac.jp)
†† Ichiro KOBAYASHI(koba@is.ocha.ac.jp)
Advanced Sciences, Graduated School of Humanities and Sciences, Ochanomizu University (†)
2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan

4.2 実験結果

HDP-LDA により推定されたトピックを表 1 に示す．全文に対して表 1 で抽出されたトピックで重み付けをし，

各トピック毎の文数をグラフで表し，式 (2) を用いて抽出したキーワードを入れたものを図 1 に，また，政治のトピックにおいて「国会」と「小泉」の出現度と「政治」のトピックの関連を表したものを図 2 に示す．

表 1: 抽出された各トピックの上位単語

トピック	上位単語	トピック名
topic0	米国 テロ 米 支援 攻撃 多発 同時 日本	政府の動向
topic1	ニューヨーク テロ 米国 ビル 世界 貿易 米	ニューヨーク
topic2	米国 テロ イスラム 攻撃 戦争 米 報復 世界	アフガニスタン
topic3	基地 米 出港 同時 予定 キティ テロ ホーク	米軍 基地
topic4	活動 実施 措置 自衛隊 規定 武器 十 支援	自衛隊 支援
topic5	預金 外貨 運用 月末 証券 残高 ドル 増加	為替 金融
topic6	スイス 航空 経営 エア 株式 赤軍 保有 灯油	航空 会社
topic7	判事 スコットランド 発効 リビア 法廷 虐殺	裁判

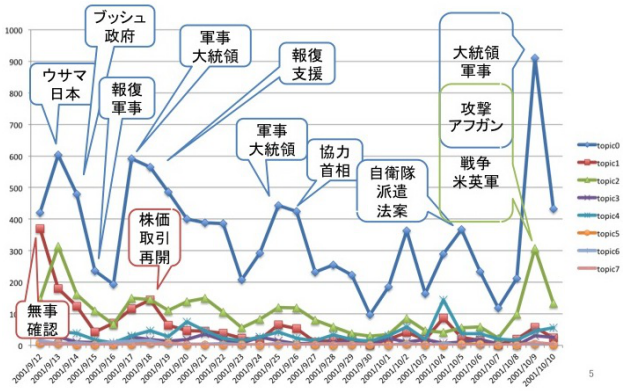

図 1: 各トピックの文数とキーワードの関係

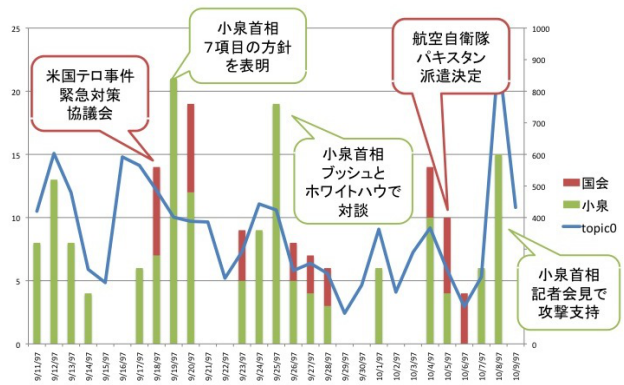

図 2: 登場人物との関わり

4.3 考察

図 1 から，対象とする期間中，政府の動向とアフガニスタンに関するトピックが多くを占めていることがわかった．10 月 9 日で topic0 と topic3 のキーワードをみてみると，「攻撃」，「アフガン」と共通の単語があり，どちらのトピックもアメリカが報復攻撃を開始したことを表している．この日において，対象とした期間の中でこれらトピックに関して，報道されていることがわかり，他のトピックと比べて非常に注目されていることがわかる．この二つのトピックの文数のグラフが連動しているところも多いことから，政治の動きとアフガニスタンの動きは大きく関わっていることがわかった．topic0 の政府の動向のトピックに関して，キーワードを追っていくと，米大統領や日本政府の動きがわかった．また topic1（ニューヨーク）のトピックに関して，文書数が多くなっている日には，テロの発生や，株取引の再開など，イベントが発生していることがわかった．topic1 はテロ発生時においては頻繁に話題になっていたが，それ以降他のトピックに視点が移ったことがわかる．図 2 では，国会や，小泉首相が行動を起こす度に，出現頻度が大きくなっていることがわかる．また 9 月 26 日において図 1 のキーワード「首相」，「協力」，と図 2 の「小泉首相とブッシュの対談」が関連していることから，重要なイベントとして捉えられていることがわかった．

5 おわりに

本研究では，事象の俯瞰分析技術の開発として，潜在意味に基づきトピックを抽出し，トピックをわたる語彙の共起関係からイベントの動向および人物の関連性を捉える手法を提案した．提案手法により，対象となる期間のイベントに関連する情報を俯瞰することができたことを確認した．今後の課題としては，リアルタイムでの俯瞰分析を行い，提案手法をより有用性の高いものにしたいと考えている．

参考文献

[1] Yee Whye Teh,Michael I. Jordan,Matthew j. Beal,David M. Blei, Hierachical Dirichlet Processes, Journal of American Statistical Association,Vol.101,2004.

[2] Weiwei Cui,Shixai Liu Tan,Conglei Shi,Yanggiu Song,Zekai J. Gao,Xin Tong,and Huamin Qu TextFlow: Towards Better Understanding of Evolving Topics in Text IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS,VOL.17,NO.12,2011.

[3] Wenwen Dou,Xiaoyu Wang,Drew Skau,William Ribarsky,and Michelle X. Zhou Leadline: Interactive Visual Analysis of Text Data through Event Identification and Exploration In Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on, pages 93-102, 2012.