文書内の潜在情報に基づく事象抽出の一考察

澤村瞳　　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
{sawamura.hitomi, koba}@is.ocha.ac.jp

1　はじめに

近年，Web 上にある大量文書データの増加に伴い，膨大な情報の中から迅速に隠された構造や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．膨大な文書データを単語やフレーズに分解し，それらの関係を一定のルールに従って分析することにより，単語間の関係や時系列の変化などを抽出でき，価値のある情報を掘り出すことができる．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を俯瞰する．また，特定人物の行動なども抽出する．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

図 1 に提案手法の概要を示す．

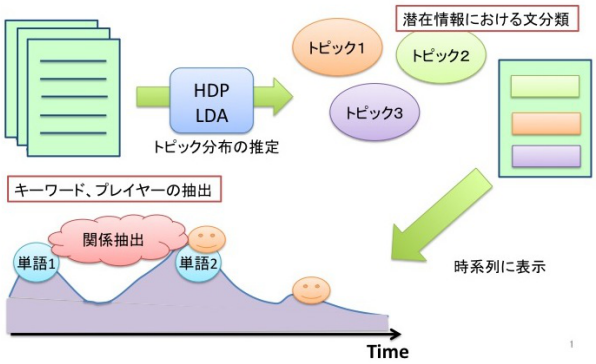

図 1: 潜在する事象の関係抽出の概要

研究の概要を処理の手順に基づいて説明する．

step 1. 潜在的意味解析
対象とする全複数文書に HDP-LDA [1] をかけ，潜在トピックを抽出する．

step 2. 文の潜在的トピック分類
文書内の各文に対して推定されたトピックの中で，一番重みの大きいトピックをその文のトピックとみなす．

step 3. キーワード，人物の抽出
step2 によりトピックを構成する重要単語や，登場人物の出現頻度を求め，各トピックおよび関連する情報を捉える．

3.2　潜在情報を考慮した文のトピック推定

3.2.1　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生成モデルであるため，LDA と異なり，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．

3.2.2 文のトピック推定

本研究では，文を代表する単語として名詞と動詞，形容詞を使用した．

また，文 i におけるトピック z に対する重要度を式 (1) に示す．ここで w は単語，ϕ をトピック内の単語の重みとする．

$$b_{z_i} = \sum_{j=1}^{V} \phi_{z_{w_j}} \cdot y_{i_{w_j}} \tag{1}$$

y は式 (2) で定義される．

$$y_{i_{w_j}} = \begin{cases} 1 & (w_j \in S_i) \\ 0 & (otherwise) \end{cases} \tag{2}$$

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 (3)，式 (4) 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 (3) 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 (4) 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．

$$Weight(w)_k^t = \frac{TF(w)_k^t}{\sum_k TF_k^t} \cdot \exp(-\lambda \times Weight(w)_k^{t-1}) \tag{3}$$

$$Weight(w)_k^t = TF(w)_k^t \cdot ISF \tag{4}$$

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．また，人物同士の Simpson 係数を用いて共起を測ることにより，人物同士の関係の強さを示した．しかし，Simpson 係数では一方の人物の出現度が低いときに，高い値が出やすい．例えば，$|X| = 1$，$|Y| = 100$，$|X \cap Y| = 1$ の場合，Simpson 係数は 1 である．そこで，閾値を設けた Simpson 係数，式 (5) を用いる．

$$R(X,Y) = \begin{cases} \dfrac{|X \cap Y|}{min(|X|,|Y|)} & (if |X| > k \wedge |Y| > k) \\ 0 \quad (otherwise) \end{cases} \tag{5}$$

ニュース記事 1 つに対して，人物 X と人物 Y が出現していれば，$|X \cap Y| = 1$ とし，加算していく．本研究では，形態素解析により「人物」,「組織」,「地域」のタグが付加されたものを人物とみなし，さらに人手で重要と判断される人物を選別した．

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し，269 個の単語を人物とみなした．HDP-LDA の設定として，サンプリング手法に Gibbs サンプリング，イテレーション 10 回，トピック分布のハイパーパラメータ $\alpha = 0.90$ と設定する．γ はガンマ分布 (1,1) より得られ，$\gamma = 0.42$ と設定する．

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 (4) を用いて抽出したキーワードを入れたものを図 2 に，また，「タリバン」と「イスラエル」の出現度と「政治の動向」と「アフガニスタン」のトピックを表したものを図 3 に，政治のトピックにおいて「国会」と「小泉」の出現度と「政治の動向」のトピックの関連を表したものを図 4 に示す．式 (5) を用いて，全文書中の人物同士の関係の強さを抽出したものを表 2 に示す．また，表 2 で登場する人物の簡単な説明を表 3 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

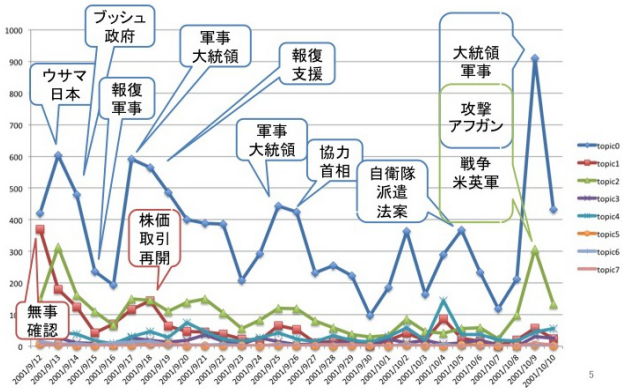

図 2: 各トピックの文数とキーワードの関係

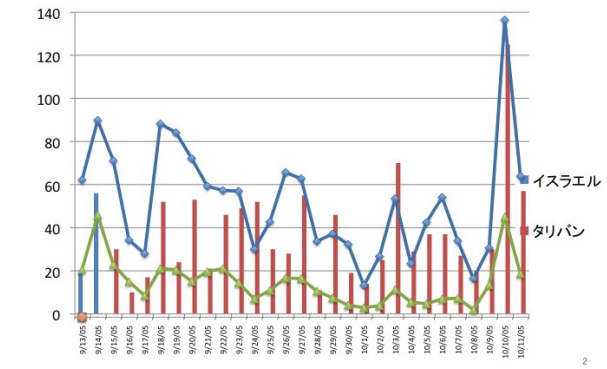

図 3: 「タリバン」と「イスラエル」の出現度

4.3 考察

図 2 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「政治の動向」と topic3「アフガニスタン」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 3 では，はじめは「タリバン」という組織を特定出来ていなかったため，タリバンは全く出現しなかったが，特定後頻繁に出現し，出現数から政治やアフガニスタンのトピックと関係してることがわかった．

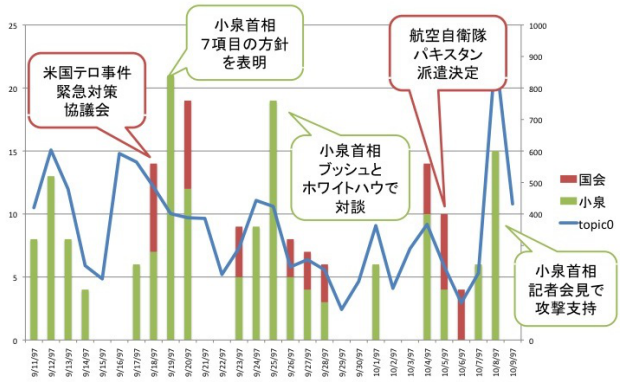

図 4: 政治のトピックに関する人物の出現度

表 2: 人物同士の関係が強い上位 10 組

順位	人物 1	人物 2	Simpson 係数
1	サミュル・ハク党首	タリバン	1.0
1	ファズド・ラフマン	タリバン	1.0
1	ブッシュ	ライス大統領補佐官	1.0
1	マスード将軍	タリバン	1.0
5	ウサマ・ビンラディン	サミュル・ハク党首	0.83
6	ブッシュ	パウエル国務長官	0.73
7	サッタル外相	タリバン	0.72
8	パウエル国務長官	タリバン	0.71
9	ブッシュ	シラク	0.70
10	ウサマ・ビンラディン	ブッシュ	0.69

図 4 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 2 のキーワード「首相」,「協力」，と図 4 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．

また，人物同士の関連性を抽出したところ,「タリバン」に着目するとイスラム組織の人物らが上位に出現し．続いて「パウエル国務長官」や「サッタル外相」が出現した．これは,「タリバン」の動きに関する対策に関する共起だと考えられる．次に,「ブッシュ」に着目すると「ライス大統領補佐官」や「パウエル国務長官」との共起はアメリカの政治に関することであり,「シラク」は外交を示していることがわかる．これらのことから，同じ組織や国に所属している組み合わせが上位に出てくることがわかった．

表 3: 人物の詳細

人物	詳細
サミュル・ハク党首	パキスタンのイスラム原理主義政党「イスラム聖職者協会」の党首，ウサマ・ビンラディンと親密な関係
ファズド・ラフマン	パキスタンのイスラム武装組織「ハルカト・ムジャヒディン」の代表「ハルカト」は「アルカイダ」の関連組織
タリバン	パキスタンとアフガニスタンで活動するイスラム主義の団体
ブッシュ	アメリカ大統領
マスード将軍	9 月 9 日に自爆テロにより暗殺された，アフガニスタンイスラム国国防大臣
ライス大統領補佐官	アメリカ大統領補佐官
ウサマ・ビンラディン	アルカイダの司令塔，911 テロの首謀者
パウエル国務長官	アメリカ国務長官
サッタル外相	パキスタンの外相，パキスタンはタリバン運動が盛である
シラク	フランス大統領

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.