複数の粒度でのLDA適用結果におけるトピック集約*

井上 祐輔† 小池 大地‡ 宇津呂 武仁‡ 神門 典子§
筑波大学理工学群工学システム学類† 筑波大学大学院 システム情報工学研究科‡
国立情報学研究所§

1 はじめに

トピックモデルの一種である潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation) [2] においては，入力として，文書集合とトピック数 K を与えることにより，トピック z_n における語の分布 $P(w|z_n)$ と，文書 d におけるトピックの分布 $P(z_n|d)$ が推定される．ここで，本論文では，複数の粒度での LDA 適用結果におけるトピックの冗長性と関連性に着目する．通常，LDA においては，トピック数が少なすぎる場合には，推定されたトピックにおける話題の多様性は低い．一方，トピック数が多すぎる場合には，トピック数が少ない場合には現れなかった新たな話題を示すトピックが存在する一方で，冗長なトピックも多数出現する．文書集合の効率のよい俯瞰というタスクを想定する場合，話題の多様性をなるべく大きくすることと裏表の関係にあることとして，複数の話題の間の関連性を考慮して，冗長な話題を集約するとともに，関連する話題の対応付けを最大限行なった状態で文書集合を閲覧する技術が不可欠である．そこで，本論文では，複数の粒度での LDA 適用結果において，冗長なトピックを集約しつつ，関連するトピックを対応付けて示すことにより，文書集合におけるよりきめ細かなトピック分布を提示する枠組みを提案する．

図 1 に示す本論文の枠組みにおいては，トピック数が少ない場合の LDA 適用結果とトピック数が多い場合の LDA 適用結果の間でトピックの間の対応付けを行い，それらのトピックを，(1) 一つのトピックに集約される冗長なトピック，(2) 一つのトピックには集約されないが関連性の強いトピック組，(3) 他のいずれのトピックとも関連しない独立した話題のトピック，に分類してトピック分布を提示する．本論文では，この枠組みを，震災に関連するニュース記事集合に対して適用し，その有効性を示す．

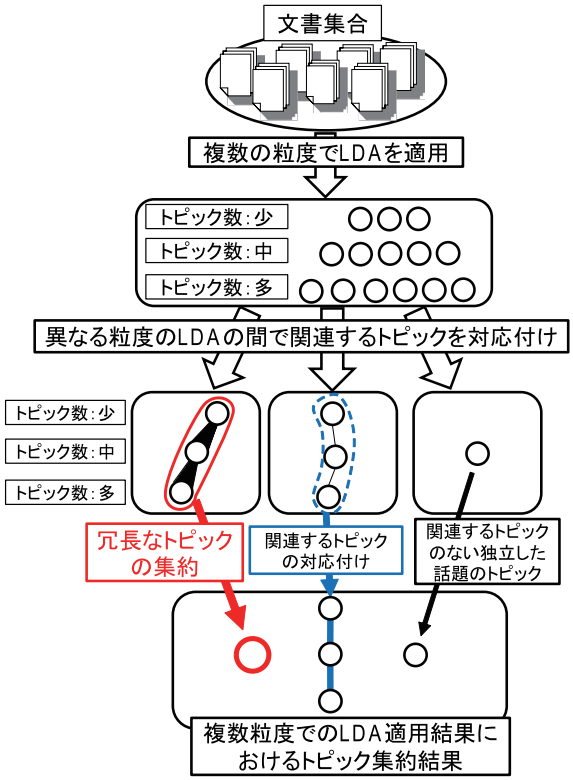

図 1: 複数の粒度での LDA 適用結果におけるトピック集約の枠組み

2 分析対象ニュース記事

分析対象ニュース記事として，2011 年 3 月 11 日から 12 月 29 日までの日付のものを，日経新聞[1]，朝日新聞[2]，読売新聞[3] の各新聞社のサイトから収集した 70,005 記事，23,237 記事，および，50,286 記事の合計 143,528 記事を用いた．その後，震災関係の 7 語[4] およびそのリダイレクトを Wikipedia から収集し，それらのうちの少なくとも一つがニュース記事中に出現するものだけを分析対象とした．その結果，各新聞社の記

*Aggregating Topics of Multi-Grain LDA
†Yusuke Inoue, College of Engineering Systems, School of Science and Engineering, University of Tsukuba
‡Daichi Koike, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba
§Noriko Kando, National Institute of Informatics

[1] http://www.nikkei.com/
[2] http://www.asahi.com/
[3] http://www.yomiuri.co.jp/
[4] 福島県，放射能，津波，東京電力，原子力発電所，放射線，原子力発電．

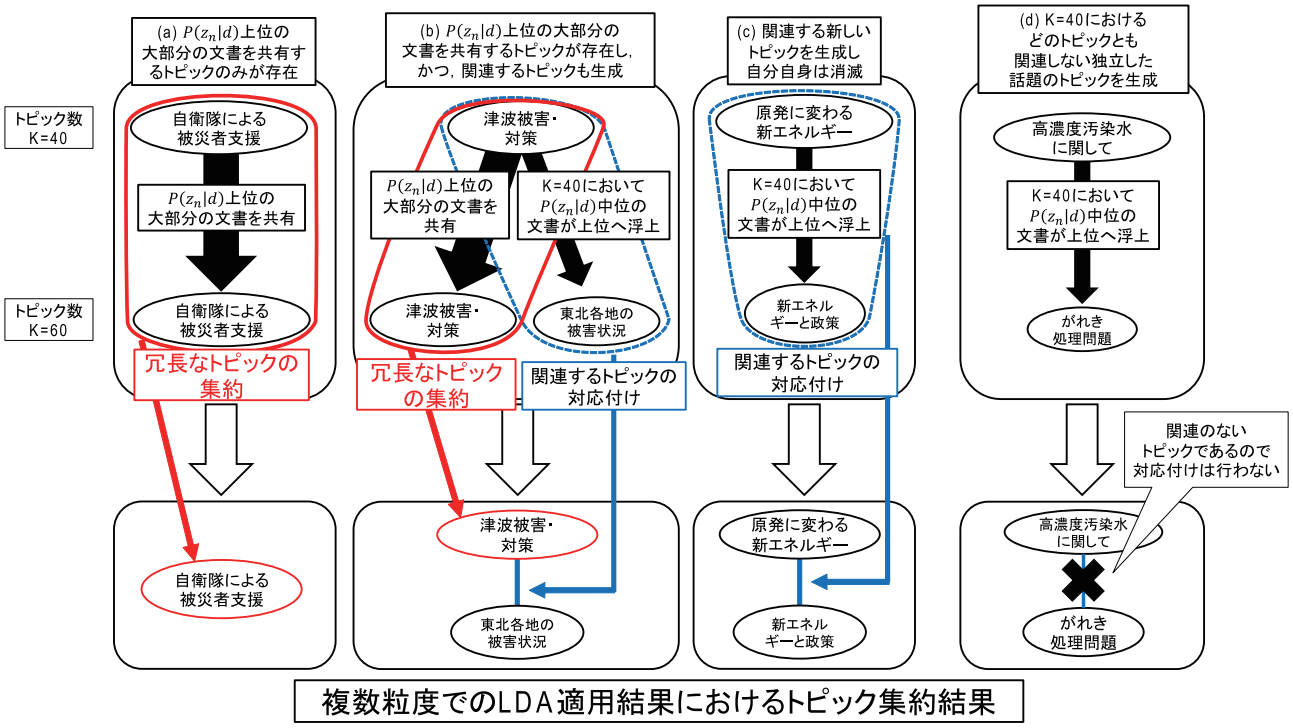

図 2: 複数の粒度での LDA 適用結果におけるトピックの対応付け・集約

事数は，日経新聞が 11,006 記事，朝日新聞が 4,988 記事，読売新聞が 8,368 記事，合計 24,458 記事となった．

3 トピックモデル

3.1 潜在的ディリクレ配分法

本論文では，トピックモデルとして潜在的ディリクレ配分法 (LDA; Latent Dirichlet Allocation) [2] を用いる．LDA を用いたトピックモデルの推定においては，語 w の列によって表現された文書の集合と，トピック数 K を入力として，各トピック z_n $(n = 1, \ldots, K)$ における語 w の確率分布 $P(w|z_n)(w \in V)$，及び，各文書 b におけるトピック z_n の確率分布 $P(z_n|b)$ $(n = 1, \ldots, K)$ を推定する．これらを推定するためのツールとしては，GibbsLDA++[5]を用いた．LDA のハイパーパラメータである α，β には，GibbsLDA++の基本設定値である $\alpha = 50/K$，$\beta = 0.1$ を用い，Gibbs サンプリングの反復回数は 2,000 とした．

3.2 文書に対するトピックの割り当て

本論文では，一つのニュース記事に対して，トピックを一意に割り当てる．文書集合を D，トピック数を K，1 つの文書を $d(d \in D)$ とすると，トピック $z_n(n = 1, \ldots, K)$ の記事集合 $D(z_n)$ は以下の式で表される．

$$D(z_n) = \left\{ d \in D \;\middle|\; z_n = \underset{z_u (u=1,\ldots,K)}{\operatorname{argmax}} P(z_u|d) \right\}$$

ここでは，文書 d におけるトピックの分布において，確率が最大のトピックを文書 d に割り当てる．

[5] http://gibbslda.sourceforge.net/

4 複数の粒度での LDA 適用結果におけるトピック集約の枠組み

異なる粒度の LDA 適用結果の間でトピックの対応付けを行った結果においては，図 2 に示す (a)~(d) の四通りの場合が存在する[6]．

(a) $K = 40$ と $K = 60$ の間で，$P(z_n|d)$ 上位 10 文書のうちの 5 文書以上を共有するトピック組 $\langle z_i^{40}, z_j^{60} \rangle$ が存在する場合，トピック z_i^{40} と z_j^{60} は同一の話題の冗長なトピックであると判定する．さらに，$K = 40$ と $K = 60$ の間で，z_i^{40}，z_j^{60} と話題の関連するトピックが他に存在しない場合．

(b) $K = 40$ と $K = 60$ の間で，$P(z_n|d)$ 上位 10 文書のうちの 5 文書以上を共有するトピック組 $\langle z_i^{40}, z_j^{60} \rangle$ が存在する場合，トピック z_i^{40} と z_j^{60} は同一の話題の冗長なトピックであると判定する．さらに，トピック z_i^{40} における $P(z_n|d)$ 中位の文書が，$K = 60$ における別のトピック z_k^{60} における $P(z_n|d)$ 上位へ浮上する場合で，かつ，トピック z_i^{40} と z_k^{60} は，話題は関連するが別のトピックであるとして対応付けられる場合．

[6] トピック数として，$K = 40$ の場合のトピックモデルと $K = 60$ の場合のトピックモデルの対応付けを行った．各トピックにおいて，$P(z_n|d)$ の降順で上位の 20 文書を対象として各トピックの話題の分析を行った．上位 20 文書の話題がまとまっていないトピック，および，震災とは無関係の話題のトピック ($K = 40$ の場合，11 トピック．$K = 60$ の場合，13 トピック) については，ノイズトピックとして扱い，トピック対応付けを行わなかった．

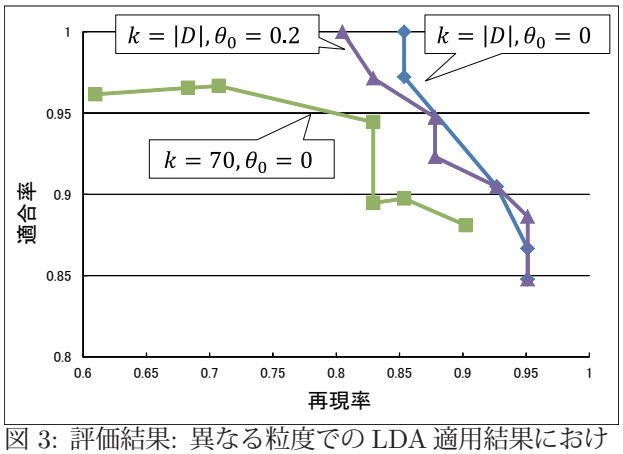

図 3: 評価結果: 異なる粒度での LDA 適用結果におけるトピックの対応付け

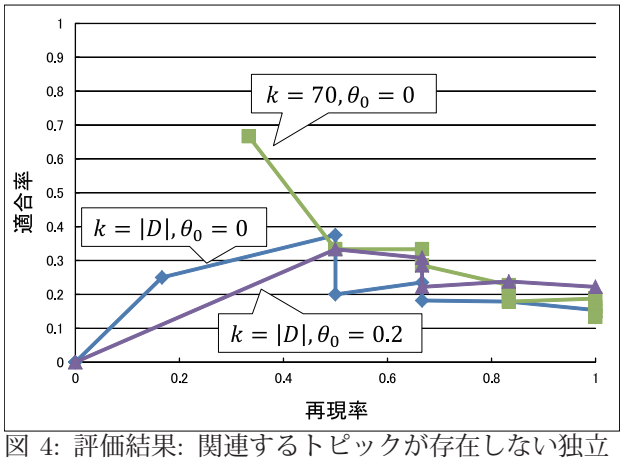

図 4: 評価結果: 関連するトピックが存在しない独立した話題のトピックの同定

(c) $K = 40$ におけるトピック z_i^{40} における $P(z_n|d)$ 中位の文書が，$K = 60$ におけるトピック z_j^{60} における $P(z_n|d)$ 上位へ浮上する場合で，かつ，トピック z_i^{40} と z_j^{60} は，話題は関連するが別のトピックである場合，これらのトピックの対応付けを行う．さらに，$K = 60$ において，トピック z_i^{40} と話題が同一であるトピックが存在しない場合，トピック z_i^{40} 自身は，$K = 60$ において消滅したと判定する．

(d) $K = 40$ におけるトピック z_i^{40} における $P(z_n|d)$ 中位の文書が，$K = 60$ におけるトピック z_j^{60} における $P(z_n|d)$ 上位へ浮上するが，トピック z_i^{40} と z_j^{60} は，話題が関連しない独立な話題のトピックの場合．

5 異なる粒度での LDA 適用結果におけるトピックの対応付け

5.1 対応付け手順

まず，トピック数 $K_1 < K_2$ として，トピック数 K_1 における LDA 適用結果におけるトピックの集合を Z_1，$z_1 \in Z_1$ とし，トピック数 K_2 における LDA 適用結果におけるトピックの集合を Z_2，$z_2 \in Z_2$ とする．本論文では，各トピック z_n における文書集合 $D(z_n)$ 中において，$P(z_n|d)$ 上位 k 個に含まれ，$P(z_n|d) \geq \theta_0$ となる文書集合として次式を定義し[7]，

$$D(z_n, k, \theta_0) = \Big\{ d \in D(z_n) \;\Big|\; d \text{ は } P(z_n|d) \text{ の降順で上位の } k \text{ 個に含まれ，かつ，} P(z_n|d) \geq \theta_0 \Big\}$$

この集合間の dice 係数

$$\mathrm{dice}(D(z_1, k, \theta_0), D(z_2, k, \theta_0)) = \frac{2 \times \big|D(z_1, k, \theta_0) \cap D(z_2, k, \theta_0)\big|}{\big|D(z_1, k, \theta_0)\big| + \big|D(z_2, k, \theta_0)\big|}$$

によって，トピック z_1 と z_2 の間の類似度を測定する．そして，この dice 係数の下限 lb を満たすトピック組の集合 $ZZ(Z_1, Z_2, lb)$

$$ZZ(Z_1, Z_2, lb) = \Big\{ \langle z_1, z_2 \rangle \;\Big|\; z_2 \in Z_2,\; z_1 = \operatorname*{argmax}_{z_1' \in Z_1} \mathrm{dice}(D(z_1', k, \theta_0), D(z_2, k, \theta_0)),\; \mathrm{dice}(D(z_1, k, \theta_0), D(z_2, k, \theta_0)) \geq lb \Big\}$$

によって，異なるトピック数において推定されたトピックモデルにおけるトピックの対応付け結果を表現する．また，この対応付けの際にどのトピックにも対応付けられなかったトピックの集合 $Z_2^{\phi}(Z_1, Z_2, lb)$ を次式で定義する．

$$Z_2^{\phi}(Z_1, Z_2, lb) = \Big\{ z_2 \in Z_2 \;\Big|\; \forall z_1 \in Z_1, \langle z_1, z_2 \rangle \notin ZZ(Z_1, Z_2, lb) \Big\}$$

5.2 評価

5.2.1 異なる粒度での LDA 適用結果におけるトピックの対応付けの評価

評価の際には，人手によって対応付けをおこなったトピック組の参照用集合 $ZZ_r(Z_1, Z_2)$ を用いて，以下の再現率，適合率によって評価を行う．

$$\text{再現率} = \frac{\big|ZZ(Z_1, Z_2, lb) \cap ZZ_r(Z_1, Z_2)\big|}{\big|ZZ_r(Z_1, Z_2)\big|}$$

$$\text{適合率} = \frac{\big|ZZ(Z_1, Z_2, lb) \cap ZZ_r(Z_1, Z_2)\big|}{\big|ZZ(Z_1, Z_2, lb)\big|}$$

$K_1 = 40$，$K_2 = 60$ として，$D(z_n, k, \theta_0)$ におけるパラメータ k および θ_0 のいくつかの組合わせのうち，トピック対応付け性能が高かったものについて，dice

[7] トピック z_n を割り当てられた全文書集合 $D(z_n)$ に対応する集合は，$D(z_n,\ k = |D|,\ \theta_0 = 0)$ となる．

表 1: トピック数 $K = 40$ と $K = 60$ の間で対応関係にあるトピックの組 (抜粋)

<table>
<tr><td rowspan="2">対応付けの際のトピック間の関係</td><td colspan="2">トピック数</td><td colspan="3">各トピックの話題 (例)</td></tr>
<tr><td>40</td><td>60</td><td>$K = 40$ (29 トピックを分析)</td><td colspan="2">$K = 60$ (47 トピックを分析)</td></tr>
<tr><td rowspan="3">(a) $P(z_n|d)$ 上位の大部分の文書を共有するトピックのみが存在</td><td rowspan="3">7</td><td colspan="3">九州電力やらせメール問題</td></tr>
<tr><td colspan="3">学校，子供たちへの影響</td></tr>
<tr><td colspan="3">自衛隊による被災地支援</td></tr>
<tr><td rowspan="2">(b) $P(z_n|d)$ 上位の大部分の文書を共有するトピックが存在し，かつ，関連する新しいトピックも生成</td><td rowspan="2">1</td><td rowspan="2">2</td><td colspan="3">津波による被害・対策について</td></tr>
<tr><td>—</td><td colspan="2">東北各地の被害状況</td></tr>
<tr><td rowspan="2">(c) 関連する新しいトピックを生成し，自分自身は消滅</td><td rowspan="2">21</td><td rowspan="2">32</td><td>原発に変わる新エネルギーについて</td><td>新エネルギーとエネルギー政策</td><td>脱原発とエネルギー政策</td></tr>
<tr><td>原発建設稼働計画への影響</td><td>新規原発建設への影響</td><td>中部電力浜岡原発</td></tr>
<tr><td rowspan="3">(d) K = 40 におけるどのトピックとも関連しない独立した話題のトピックを生成</td><td rowspan="3">—</td><td rowspan="3">6</td><td rowspan="3">—</td><td colspan="2">海水注入中断問題について</td></tr>
<tr><td colspan="2">東電の株主総会，経営責任問題</td></tr>
<tr><td colspan="2">がれきの処理について</td></tr>
</table>

係数の下限 lb を変化させて再現率・適合率の推移をプロットした結果を図 3 に示す．また，トピックの具体例を表 1 に示す．この場合，全文書集合 $D(z_n) = D(z_n, k = |D|, \theta_0 = 0)$ の場合に最も高い性能となった．

5.2.2 関連するトピックが存在しない独立した話題のトピックの判定の評価

評価の際には，人手によって作成した参照用集合 $Z_2^{\phi_r}(Z_1, Z_2)$ を用いて，以下の再現率，適合率によって評価を行う．

$$\text{再現率} = \frac{\left|Z_2^{\phi}(Z_1, Z_2, lb) \cap Z_2^{\phi_r}(Z_1, Z_2)\right|}{\left|Z_2^{\phi_r}(Z_1, Z_2)\right|}$$

$$\text{適合率} = \frac{\left|Z_2^{\phi}(Z_1, Z_2, lb) \cap Z_2^{\phi_r}(Z_1, Z_2)\right|}{\left|Z_2^{\phi}(Z_1, Z_2, lb)\right|}$$

前節と同様に，再現率・適合率の推移をプロットした結果を図 4 に示す．また，トピックの具体例を表 1 に示す．この場合，前節とは逆に，文書集合 $D(z_n, k = 70, \theta_0 = 0)$ の場合に最も高い性能となった．

6 関連研究

文献 [4] においては，本論文と同様に，できるだけ冗長性を排して文書集合をクラスタリングするタスクを集合被覆問題として定式化している．対象文書集合と一般的な文書集合における単語の出現確率の差に基づいて，話題ラベルとなる n グラムを抽出し，貪欲法によって集合被覆問題を解く手法を提案している．また，文献 [1,6] では，検索された個々の Web ページに対してラベルの付与を行い，付与されたラベルに基づいて分類を行う手法を提案している．以上の研究では，トピックモデル以外の方式に基づいて，文集集合中の内容のまとまりを同定するという手法が用いられている．

一方，文献 [3] においては，トピックモデルの一種である PLSI(Probabilistic Latent Semantic Indexing) を用いて，Web ページの検索結果をトピックへと分類し，AIC に基づいて，3~5 のトピック数の範囲で最適なトピック数を決定し，各トピックに対して要約文を付与するという手法を提案している．また，文献 [5] においては，文書集合に対して LDA を適用し，トピックの特徴語とその特徴量をベクトルで表し，余弦類似度を用いてトピック間の類似度を計算することで，LDA における適切なトピック数を自動的に推定する手法を提案している．この研究では,「尖閣諸島問題」という非常に限定された話題の，80 程度の記事を対象にしているが，本論文では,「東日本大震災」という比較的広い内容の，約 24,000 の大規模な記事を対象としており，記事数の規模の点において大きく異なっている．

7 おわりに

本論文では，複数の粒度での LDA 適用結果におけるトピックの冗長性と関連性に着目した．そして，複数の粒度での LDA 適用結果において，冗長なトピックを集約しつつ，関連するトピックを対応付けて示すことにより，文書集合におけるよりきめ細かなトピック分布を提示する枠組みを提案し，その有効性を示した．

参考文献

[1] 馬場康夫, 黒橋禎夫. キーワード蒸留型クラスタリングによる大規模ウェブ情報の俯瞰. 情報処理学会論文誌, Vol. 50, No. 4, pp. 1399–1409, 2009.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[3] 原島純, 黒橋禎夫. PLSI を用いたウェブ検索結果の要約. 言語処理学会第 16 回年次大会論文集, pp. 118–121, 2010.

[4] P. Muthukrishnan, J. Gerrish, and D. R. Radev. Detecting multiple facets of an event using graph-based unsupervised methods. In *Proc. 22nd COLING*, pp. 609–616, 2008.

[5] 芹澤翠, 小林一郎. 文書内のトピック数を考慮したトピック追跡の試み. 言語処理学会第 18 回年次大会論文集, pp. 1196–1199, 2012.

[6] 戸田浩之, 中渡瀬秀一, 片岡良治. 特徴的な固有表現を用いたラベル指向ナビゲーション手法の提案. 情報処理学会論文誌：データベース, Vol. 46, No. SIG 13(TOD 27), pp. 40–52, 2005.