

論文

卒業研究要旨を活用したラーニング・アナリティクス —中等教育学校における卒業研究要旨31年分の定量的記述—

黒宮 寛之^{1,a)} 日高 一郎¹ 沖濱 真治² 山本 義春¹

概要: 本研究は校内データベース蓄積された生徒の卒業研究要旨 31 年分を対象とし、自然言語処理を用いてその経年変化を定量的に視覚化した試みである。解析対象は 1984 年から 2014 年に都内中等教育学校に提出された卒業研究要旨 3,380 編で、それぞれの要旨について要旨の長さとしリーダビリティ、およびトピックの配分量を計算した。分析の結果、近年の要旨は取り組み初期の頃の要旨と比較して要旨の長さが増加し、日本語としての難易度が高くなっていることが分かった。また要旨のトピックについては「調査」や「分析」など実証的な研究に関するトピックの成長率が大きかった。これらの結果は調査対象校の探究的な学習の性質を的確に捉えており、他の学校においても校内データベースの整備が進む中で、このような解析は多くの学校で指導の振り返りとして有用であろう。

キーワード: ラーニング・アナリティクス, 自然言語処理, トピックモデル, 探究的な学習

Learning Analytics for Graduation Work: Quantitative Description about Graduation Work's Abstract over 31 Years at a Secondary School

HIROYUKI KUROMIYA^{1,a)} ICHIRO HIDAKA¹ SHINJI OKIHAMA² YOSHIHARU YAMAMOTO¹

Abstract: The aim of the study is to illustrate secular changes of graduation work on a high school using Natural Language Processing (NLP). We analyzed 3,381 pieces of paper's abstract over 31 years (from 1984 to 2014) which is written by students. We calculated the length, the readability and topic distributions of the paper. We used Latent Dirichlet Allocation (LDA) to estimate topic allocation of it. From the results, the length got longer and the readability got lower with the flow of time. The topic related to "investigation" and "analysis" show high growth rate among 40 topics. They reveal important aspects of research-based active learning in that school. Since school education is increasingly migrating to information technology, this kind of analysis can be good way to review your teaching style at a school.

Keywords: Learning Analytics, Natural Language Processing, Topic Model, Research Based Active Learning

1. はじめに

1.1 研究の背景

2014 年に次期学習指導要領案としてアクティブラーニン

グが取り入れられて以降 [15], 卒業研究に代表されるような生徒の興味関心に基づいて主体的にテーマを研究する探究的な学習は多くの学校で導入が検討されている [17]. しかし現在探究的な学習に関する分析は主に大学の研究者や現場の教員による定性的な研究の一環として行われることが多い [12].

一方で近年特に高等教育において教育の情報化が進み様々な学習データがログとして取得できるようになった現

¹ 東京大学大学院教育学研究科
Graduate School of Education, University of Tokyo

² 東京大学教育学部附属中等教育学校
The University of Tokyo Secondary School

a) kuromiya@p.u-tokyo.ac.jp

状を受けて、「ラーニングアナリティクス」と呼ばれる概念が注目を集めている [10]. ラーニング・アナリティクスとは「学習とその環境を理解し最適化するための、学習者とその文脈に関するデータの測定・収集・分析・報告」であり [6], 現在大規模オンライン学習システム (MOOC) や学習管理システム (LMS) を中心に研究が進められているが, 実際の現場で探究的な学習を対象にした研究は少ない.

1.2 本研究の目的

本研究の目的は探究的な学習の成果物としての卒業研究要旨の 31 年間の経年的な変化を定量的な形で可視化し, 特徴的な変化を抽出することである. 実体として捉えるのが難しい探究的な学習の指導の結果を視覚的・定量的な形で提示することによって, 現場の教員及び教育学の研究者にとって効率的な指導の振り返りが可能になると考えられる.

1.3 本研究の立ち位置

本研究は既存のラーニング・アナリティクス研究と比較して, 以下の 2 点に特徴がある.

1 点目は中等教育機関を対象としていることである. ラーニング・アナリティクスの研究は主に, 在籍者が多く学習管理システムが整備されている高等教育機関を対象に行われることが多い. 一方で近年高大接続や主体的・対話的で深い学びの検証の観点から初等中等教育においてもラーニング・アナリティクスの実施が求められている [13]. 本研究は中等教育機関における探究的な学習を対象にした学習データの分析を行っている点で希少性がある.

2 点目は本研究が生徒の学習の成果物そのものを対象としていることである. 従来のラーニング・アナリティクス研究では分析に適した形式で分析用にデータを取り直すものが多かった. 代表的な例が授業後のアンケートを材料としたラーニングアナリティクス研究であるが, この方法では生徒に余分な負担を強いることになり, またデータのリアリティが落ちるといった懸念もある. 本研究では自然言語処理技術を用いることでレポートやプレゼンなど非定型なデータになりがちな, 探究的な学びの成果物に対する定量的な分析を可能にした.

1.4 調査対象校における卒業研究の取り組みについて

本研究の対象は都内公立中高一貫校 A 校 (以下, 調査対象校) における卒業研究の取り組みである. 調査対象校では 1983 年から「卒業研究」と題して, 総合的な学習の時間を使い生徒に探究的な学びの指導を行ってきた. 調査対象校の定義によれば, 卒業研究とは生徒ひとりひとりが, 興味関心を持っている独自のテーマを準備し, これまでに培った自主的学習態度を十分に活用して, 指導教官の助言を得ながら, 長期間 (実質 1 年半) にわたって考察・調査・研究を行い, 結果をまとめ上げるものであるとされている.

2. 方法

2.1 解析対象

解析対象は 1984 年～2014 年の間に調査対象校に提出された卒業研究要旨 3,380 本 (総文字数 4,054,823 字) である. 調査対象校では 1 学年あたり 100 人強の生徒が在籍しており, 各学年のデータが 31 年分あるため, これは 31 年間の卒業生をほぼ不足なく取得できたことを意味している. ただし解析に使用する指標の都合上英語と日本語が混在した文書 (英訳を掲載していた要旨 1 本, 洋楽の歌詞の大部分を引用していた要旨 1 本, プログラムコードを掲載していた要旨 1 本) については英語で書かれた箇所を除いて解析に使用した.

2.2 計算指標

本研究では要旨の定量化にあたって以下の 3 つの指標を使用した.

2.2.1 要旨の文字数

本研究における要旨の文字数はタイトル・記号を含めて要旨の文字数をカウントしたものである. 現在卒業研究要旨には 1520 字以上 1600 字以下という制限があるため意味のある変化が観察されない可能性もあったが, 要旨の文字数は最も基礎的かつ簡便な統計量であるため今回の解析に使用した.

2.2.2 jReadability

多くの小論文自動採点システムで文章の難易度に関する指標が採用されているように [3], 作文指導において文章のリーダビリティは重要な概念である. そのため本研究では文章の難易度を表す指標として李 (2017) [18] の提唱した jReadability を見ることとした. これは日本語能力試験を基準として開発された指標であり, スコアが低いほど日本語として難易度の高い文章であることを意味する. 算出式は式 1 で表される.

$$\begin{aligned} jReadability = & \\ & (\text{平均文長} \times -0.056) + (\text{漢語率} \times -0.126) \\ & + (\text{和語率} \times -0.042) + (\text{動詞率} \times -0.145) \\ & + (\text{助詞率} \times -0.044) + 11.724 \end{aligned} \quad (1)$$

李 (2017) によれば jReadability 2.5～4.4 の文書は日本語中級, それより小さい文書は日本語上級, 大きい文書は日本語初級と判断される. なお, 文章の単語単位への切り分けについてはコンピュータによる形態素解析を使用し, 形態素解析器として MeCab (ver. 0.996) [5], 形態素解析辞書として Unidic (ver. 2.1.2) [11] を使用した.

2.2.3 潜在トピック配分量

文体の他に要旨の内容を表す指標として, 本研究では代表的なトピックモデルの一つである Latent Dirichlet Allo-

cation (LDA)[1] を使用した。トピックモデルとは文書の集合が与えられたときに特定の文書に特定の話題（トピック）がどの程度含まれているかを推定するものである [9]。LDA では文書は複数のトピックから構成されると仮定する。具体的には、トピック z_{d_n} がパラメータ α のディリクレ分布から、単語 w_{d_n} がトピック z_{d_n} 毎に異なるディリクレ分布（パラメータ β ）によって生成されるとして、与えられた文書集合 D からトピックの単語分布と文書のトピック分布の両方を推定する。式 2 に LDA の尤度を示す。パラメータ推定にあたって Blei et al.,(2003)[1] では EM アルゴリズムによる推定が紹介されていたが、今回はマルコフ連鎖モンテカルロ法の一つであるギブスサンプリングを適用した。

$$p(D|\alpha, \beta) = \prod_{d=1}^M \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{d_n}} p(z_{d_n}|\theta_d) p(w_{d_n}|z_{d_n}, \beta) \right) d\theta_d \quad (2)$$

LDA の実装にあたっては、テキストを単語単位に分割し名詞を抜き出した上で、R の topicmodels パッケージ (ver.0.2.6) を用いた。LDA のハイパーパラメータは、method = "gibbs", K = 40, $\alpha=50/40$, $\delta = 0.1$, iter=2000, burnin=0 とした。ただし K はトピック数、 α, δ はそれぞれトピック分布、単語分布の事前分布のパラメータである。ほとんどのパラメータはパッケージのデフォルトの値を用いた [4] が、トピック数の決定にあたっては R の ldatuning パッケージ (ver.0.2.0)[8] を利用し、トピック間のコヒーレンスが最大となるトピック数を選択した。これは Deveaud et al.,(2015)[2] によって提案された指標で、トピック間のすべての組み合わせについて KL ダイバージェンスを計算し、それらの平均を比較することによって最適なトピック数を決定する (式 3)。一般にトピック数の選択にはモデルの予測精度を測る perplexity を使用した指標が使われることが多いが、今回はトピックの解釈可能性を重視し、予測精度ではなくトピックの品質を測定する Ad-hoc な指標を用いて選択を行った。

$$\hat{K} = \arg \max_K \frac{1}{K(K-1)} \sum_{(k, k') \in T_K} D(k||k') \quad (3)$$

さらに、本研究では抽出されたトピックの中から経年変化に一定の傾向があるトピックを選んで提示するため、トピックの成長率等比数列的な成長を仮定し、経年成長率 r を以下の式によって定義した。

$$y_t = y_1(1+r)^{(t-1)} \quad (4)$$

経年成長率 r は対数変換したトピック配分量 y を目的変数、収録年 t を従属変数として線形回帰によって推定でき

る [7].

$$r_{OLS} = e^{\hat{\beta}} - 1 \quad (5)$$

ただし $\hat{\beta}$ は推定された回帰係数である。

3. 結果

3.1 要旨の文字数

要旨の文字数の経年変化は図 1 のようになった。太線が平均値の推移を表しており、上下の領域は標準偏差を表している。図より 1994 年から 1995 年にかけて平均値が大きく上昇したこと、および過去の要旨ほど学年内のばらつきが大きく近年の要旨ほどばらつきが小さいことが読み取れる。

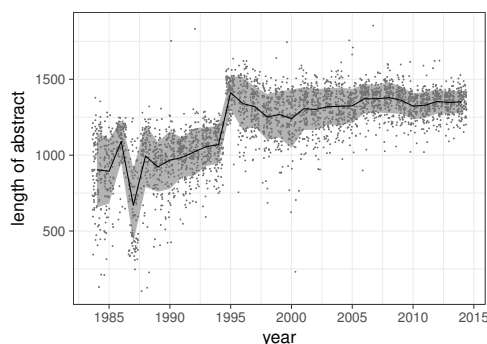


図 1 要旨の文字数の経年変化
縦軸は文字数 (字)

3.2 jReadability

jReadability の経年変化は図 2 のようになった。太線が平均値の推移を表しており、上下の領域は標準偏差を表している。図より平均値は微減傾向にあり、標準偏差は特に 1984 年～1994 年にかけて減少傾向にあることが分かる。

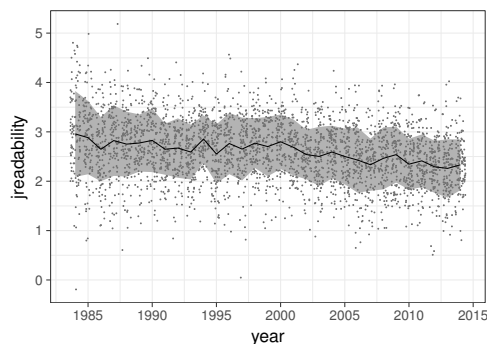


図 2 jReadability の経年変化
縦軸は jReadability スコアであり、2.5 / 4.4 を境として日本語上級 / 中級 / 初級と区分けされる。

3.3 潜在トピック配分量

まず図3にldatuningパッケージによる最適なトピック数探索の結果を示す。図より,Deveaud et al.,(2014)による指標から最適なトピック数として40が選択された。そして表1にトピック数を40としたLDAによって抽出されたトピックの内訳の一部を示す。右側の単語は当該トピックの生成確率上位4~5単語である。さらに、ここで抽出されたトピックをもとに、それらの収録年ごとの平均含有率がどのように変化しているかをみた。抽出されたトピックの経年成長率の推定結果は図4のようになった。図より、経年成長率の大きいトピックとしてトピック2(調査),16(考察),33(改善)があり、経年成長率が負のトピックとしてトピック1(卒業),28(本),34(世界)があることがわかる。実際に成長率上位と下位3つのトピックについて、経年変化を図5に示す。また抽出したすべてのトピックの内訳と経年変化を付録の表A.1と図A.1に記した。

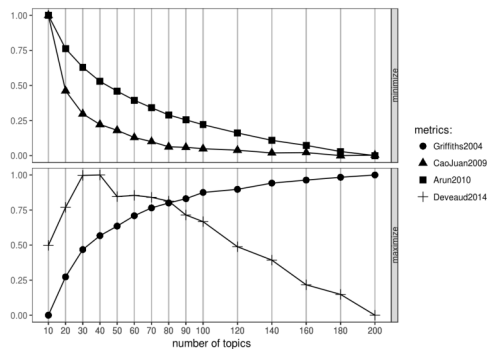


図3 トピック数を変化させたときのLDAの評価指標の推移
ldatuningパッケージはデフォルトで4つの指標を出力するが、今回着目するのは下の段に+記号で書かれたDeveaud2014のみである。横軸がトピックの数、縦軸が最大値を1としたときの coherence の相対値を示しており、トピック数30~40のときに最もトピック同士の重なりが小さいことがわかる。

表1 抽出されたトピック(一部)

トピック	上位の単語
Topic 1	テーマ 卒業 写真 先生 本当
Topic 2	調査 アンケート 結果 文献
Topic 9	活動 参加 高校 体験 国際
Topic 15	学校 教育 授業 生徒 学習
Topic 16	考察 分析 要素 結論 結果
Topic 18	存在 神 人々 宗教 死
Topic 28	本 絵 図書 興味 顔
Topic 33	改善 作成 方法 目的 提案
Topic 34	世界 戦争 国 経済 中国
Topic 37	情報 方法 インターネット 普及

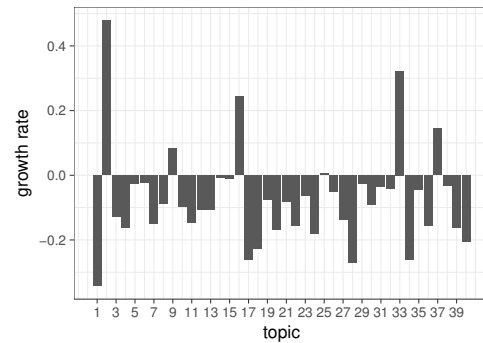


図4 推定されたトピックの成長率
横軸がトピック番号、縦軸が線形モデルによって推定された成長率を表している。

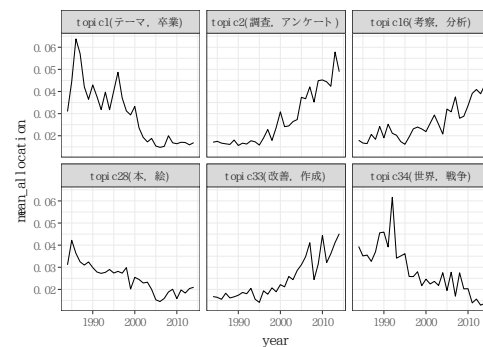


図5 平均トピック配分量の経年変化(一部)
縦軸は当該年度における全文書の平均配分率を表している。

4. 考察

4.1 変化の傾向

4.1.1 要旨の文体

図1, 図2より要旨の文字数は1995年に大きく増加し、jReadabilityは全年を通して減少傾向にあることが示された。jReadabilityが減少傾向にあるということは要旨の文章が、日本語としての難易度が高い新聞記事的な文体に近づいていることを示している。また他に、それらの学年内におけるばらつきは両者とも減少傾向にあることが示された。要旨の長さ・リーダビリティの変化の方向性については、卒業研究の成果物として研究論文的な文体が志向されることを考慮すると、教員の指導の方向性にあった望ましい変化であると言える。また、2つの指標における標準偏差の減少についても、指導の効果が特定の生徒に偏るのではなく生徒間に共有されてきているという観点で良い変化であると考えられる。

4.1.2 要旨のトピック

表1, 図4及び図5より増加傾向・減少傾向があると示されたトピックについて、以下の3つの観点からまとめる。感想文的論調について 探究的な学習について、学びの成果をレポートの形で提出させたとしても、中等教育段階では生徒の成果物はどうしても感想文的な論調にな

りがちなのではないかという懸念がある。しかし今回のデータからは「調査, アンケート」(トピック 2) や「考察, 分析」(トピック 16) など実証的な研究に関するトピックが増加傾向にあり, 反対に個人的な事柄が含まれると予測される「卒業」や「先生」に関するトピック(トピック 1) が減少傾向にあることが明らかになった。このことは卒業研究要旨において感想的な文章が減少し, 実証的傾向の強い論文が増えてきていることを示唆しており, 中等教育段階でも適切な指導を続けることで十分に探究的な学びが習得される可能性が示された。

研究テーマの選定について 「世界, 戦争」に関するトピック(トピック 34) が減少傾向にあり, 反対に「活動, 参加」に関するトピック(トピック 9) が増加傾向にあることから, 卒業研究の研究テーマの傾向として, 遠くて抽象的なテーマから身近で実地調査が可能なテーマへのシフトが起こっていることが推測される。この結果は調査対象校において配布されている卒業研究ハンドブックに「壮大なテーマは失敗しやすく身近なテーマから選ぶと良い」というアドバイスが記載されており, この記述と対応するような変化であると考えられる。

調査方法について 本データでは「情報, インターネット」を表すトピック(トピック 37) が増加傾向にあり, 反対に「本, 絵」に関するトピック(トピック 28) が減少傾向にあった。トピック内には他の単語も含まれるため一概に結論付けることはできないが, このことは卒業研究の調査方法として書籍だけでなくインターネットを活用した手法が増えてきていることを示していると考えられる。

これらの結果はこの結果は以前に調査対象校の教員に対し, 卒業研究を通して生徒に身につけて欲しい力を尋ねた研究 [14] で挙げられた「課題発見力, 批判力, 自発性, コミュニケーション能力, 分析力, 計画性, 情報収集力, 行動力, 自己肯定感」と概ね対応しており, 自然言語処理を利用した本手法の妥当性を表していると考えられる。

4.2 今後の展望

文部科学省(2016) [16] は 2020 年代における我が国の学校教育の情報化に関する文書の中で, 急速に情報化した社会の変化に対応して「授業・学習面と校務面の両面で ICT を積極的に活用し, 教育委員会・学校の取組を効果的に支援する」ことを目標としている。本研究は校内に蓄積されたデータベースを活用したものであり, 今後学校種別を問わず全国の学校で探究的な学習の指導に関連したデータが蓄積されるであろうことを考慮すると, 本研究で提案された手法は調査対象校以外の多くの学校で定量的な指導の振り返りに有効であると言える。特に本研究は生徒の成果物を対象にするため, 追加の調査を使用しない点は他校に

おいて分析を検討する上で大きな利点になるであろう。

4.3 研究限界

本研究は縦断データではなく反復横断データであり, 対象集団は各年で異なる。そのため, 本研究でみた変化が教育指導の結果によるものなのか, その年々に入学した生徒の資質的な違いによるものなのかは今回のデータからは判断できない。より詳細な指導の実態を把握するためには, 同一生徒の成果物を複数回にわたって記録し, 蓄積することが求められる。

5. 倫理申請

本研究は 2017 年 1 月 25 日東京大学ライフサイエンス委員会によって承認された(申請者: 教育学研究科 教授山本義春, 審査番号: 16 - 220)。

参考文献

- [1] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent dirichlet allocation, *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993-1022 (2003).
- [2] Deveaud, R., SanJuan, E. and Bellot, P.: Accurate and effective latent concept modeling for ad hoc information retrieval, *Document numérique*, Vol. 17, No. 1, pp. 61-84 (2014).
- [3] Dikli, S.: An overview of automated scoring of essays, *The Journal of Technology, Learning and Assessment*, Vol. 5, No. 1 (2006).
- [4] Hornik, K. and Grün, B.: topicmodels: An R package for fitting topic models, *Journal of Statistical Software*, Vol. 40, No. 13, pp. 1-30 (2011).
- [5] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying conditional random fields to Japanese morphological analysis, *Proceedings of the 2004 conference on empirical methods in natural language processing* (2004).
- [6] LAK11: About the 1st International Conference on Learning Analytics and Knowledge 2011., International Conference on Learning Analytics and Knowledge (online), available from (<https://tekri.athabascau.ca/analytics/>) (accessed 2018-05-14).
- [7] Mawson, P. et al.: Measuring economic growth in New Zealand, Technical report, New Zealand Treasury (2002).
- [8] Nikita, M.: *ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters* (2016).
- [9] 岩田具治: トピックモデル, 講談社 (2015).
- [10] 山田政寛: ラーニング・アナリティクス研究の現状と今後の方向性, 日本教育工学会論文誌, Vol. 41, No. 3, pp. 189-197 (2018).
- [11] 小木曾智信, 中村壮範: 『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システム的设计・実装・運用, 自然言語処理, Vol. 21, No. 2, pp. 301-332 (2014).
- [12] 松下佳代: ディープ・アクティブラーニング, 京都大学高等教育研究開発推進センター勁草書房 (2015).
- [13] 森本康彦, 稲垣忠: 初等中等教育におけるラーニング・アナリティクスの展望, 日本教育工学会論文誌, Vol. 41, No. 3, pp. 209-220 (2018).
- [14] 能智正博, 沖濱真治, 石橋太加志: 附属中等教育学校における卒業研究指導の現状と課題の探求, 東京大学教育

- 学研究科附属学校教育高度化センター研究紀要, p. 198 (2016).
- [15] 文部科学省：初等中等教育における教育課程の基準等の在り方について (諮問), 2014a, (http://www.mext.go.jp/b_menu/shingi/chukyo/chukyoo0/toushin/1353440.htm) 2017年, Vol. 1 (2014).
- [16] 文部科学省ほか：「2020年代に向けた教育の情報化に関する懇談会」最終まとめ, 政策特報, No. 1507, pp. 66-110 (2016).
- [17] 本田由紀：研究型アクティブラーニングの現状・課題・可能性, 東京大学大学院教育学研究科紀要, Vol. 56, pp. 245-262 (2016).
- [18] 李在鎬：日本語教育のための文章難易度に関する研究, 早稲田日本語教育学= Waseda studies in Japanese language education, No. 21, pp. 1-16 (2016).

付 録

表 A-1 LDA によって抽出されたトピック一覧
トピック番号 上位の単語

Topic 1	テーマ 卒業 写真 先生 本当
Topic 2	調査 アンケート 結果 文献 インタビュー
Topic 3	変化 現代 昔 現在 発展
Topic 4	心 人間 行動 心理 精神
Topic 5	問題 環境 解決 ゴミ 現在
Topic 6	地域 区 東京 都市 市
Topic 7	食 食品 栄養 食事 生活
Topic 8	対策 道 自転車 交通 安全
Topic 9	活動 参加 高校 体験 国際
Topic 10	知識 菓子 し 興味 実際
Topic 11	日本 アメリカ 外国 文化 国
Topic 12	事件 原因 少年 犯罪 医療
Topic 13	社会 障害 制度 裁判 国民
Topic 14	スポーツ トレーニング 運動 ゲーム 向上
Topic 15	学校 教育 授業 生徒 学習
Topic 16	考察 分析 要素 結論 結果
Topic 17	テーマ 資料 興味 内容 説明
Topic 18	存在 神 人々 宗教 死
Topic 19	音楽 音 曲 効果 歌
Topic 20	時代 歴史 文化 江戸 人々
Topic 21	製作 デザイン 技術 制作 物
Topic 22	人間 自然 動物 環境 犬
Topic 23	子ども 子供 保育 遊び 児童
Topic 24	言葉 英語 語 絵本 意味
Topic 25	比較 関係 違い 影響 差
Topic 26	企業 商品 消費 会社 農業
Topic 27	作品 映画 人物 表現 小説
Topic 28	本 絵 図書 興味 顔
Topic 29	実験 結果 数 方法 データ
Topic 30	建築 住宅 理想 空間 構造
Topic 31	駅 街 地震 商店 場所
Topic 32	色 イメージ 広告 印象 効果
Topic 33	改善 作成 方法 目的 提案
Topic 34	世界 戦争 国 経済 中国
Topic 35	高齢 介護 施設 福祉 老人
Topic 36	生活 時間 人々 日常 川
Topic 37	情報 方法 インターネット 普及 機能
Topic 38	テレビ 新聞 若者 人気 番組
Topic 39	女性 社会 差別 仕事 労働
Topic 40	機 自動 開発 地球 エネルギー

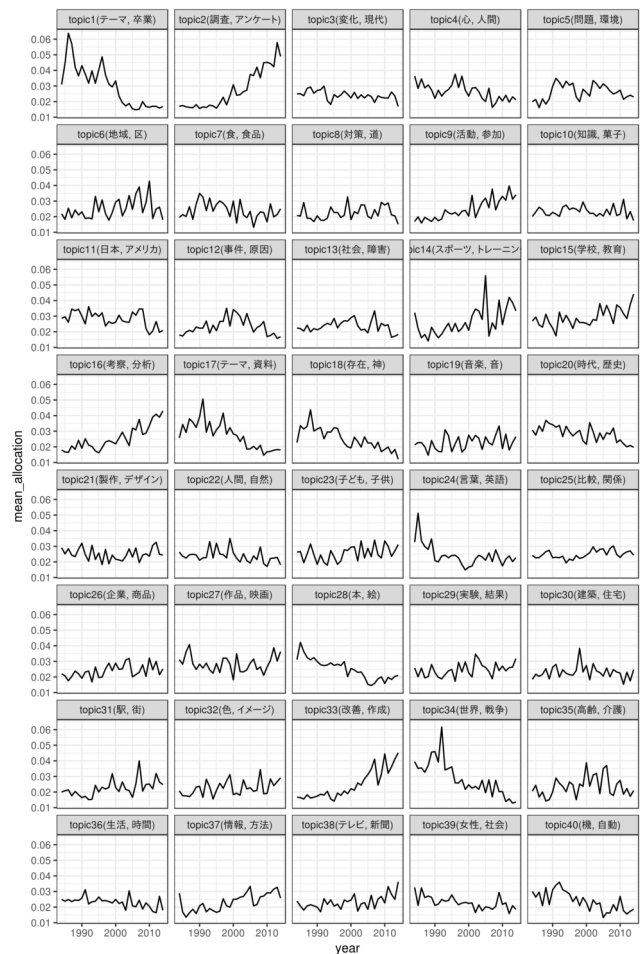


図 A-1 抽出された全トピックの平均配分率の経年変化