LDA を用いたレポート推薦システムの開発

加藤　嘉浩

電気通信大学大学院 情報システム学研究科

学位申請論文 博士 (工学)

2016 年 3 月

LDA を用いたレポート推薦システムの開発

博士論文審査委員会

主査:　植野 真臣 教授
委員:　栗原 聡 教授
委員:　大須賀 昭彦 教授
委員:　広田 光一 教授
委員:　田原 康之 准教授
委員:　川野 秀一 准教授

Reports Recommendation System Based on Latent Dirichlet Allocation

Yoshihiro Kato

abstract

We propose a reports recommender system encouraging students to learn from the others.The system can search reports that have same subject by estimating latent topics of learners' reports, and calculates distance of others' topic distributions based on Latent Dirichlet Allocation (LDA). The system recommends past others' excellent articles based on similarity of subject and contents. To be more precise, the system provides articles that has similar subject to submitted article, and has dissimilar words in an article. By recommending the reports of same subject with diverse words, beginners can improve their reports in con guration, expression and originality. In addition, we show the effectiveness of the proposed method by a subjects experiment. The proposed method fixed number of topics in LDA. For determining the number of topics, we set 1 for hyperparameters of LDA and maximize marginal likelihood. We describe some asymptotic of marginal likelihood to explain the sensitivity and hyperparameters effects. The number of topics increases monotonically as the hyperparameters increases, the number of topics monotonically decreases as it decreases. We demonstrate the efficiency of the setitng 1 for hyperparameters using simulated data and the learners' reports.

LDA を用いたレポート推薦システムの開発

加藤　嘉浩

要旨

本論文では，レポートライティングにおける他者からの学びを支援するために，過去の学ぶべきレポートを学習者に推薦するシステムを提案する．本システムの特徴は，(1) Latent Dirichlet Allocation（LDA）により，学習者のレポートの潜在的なトピックを推定し，他者レポートとのトピック分布の距離を計算して，同一の主題を扱う他者レポートを検索でき，さらに，(2) 学習者のレポートと他者レポートとの単語分布の距離を計算し，同一の主題を扱うが，内容（用いられる単語分布）の異なる評価の高い他者のレポートを多様に推薦できることである．これにより，学習者は自分と同じ主題を扱う多様な過去の優秀なレポートから，レポートライティングにおける多様なスキルを学べると期待できる．被験者実験により提案手法の有効性を示した．しかし，これまでトピック数をデータから決定する手法が確立されていなかったので，上の研究ではトピック数を決めて用いている．そこで，次に，実データからトピック数を自動的に決定する手法として，漸近解析によりハイパーパラメータが 1 .0 としたときの周辺尤度を最大化することにより，LDA のトピック数を最も正確に推定できることを提案する．本システムに組み込むことで，その有効性を示した．

目次

図目次

表目次

第 1 章

緒言

本論文では，レポートライティングにおける他者からの学びを支援するために，過去の学ぶべきレポートを学習者に推薦するシステムを提案する．他者からの学びは，単一の他者のみからよりも多様な他者からの学びの方が効果的であることが知られている．そのため，レポートライティングにおいては，他者の多様なレポートを推薦する必要がある．しかし，単に内容・表現が類似のレポートを推薦しても効果的な学習が期待できないと考える．

そこで本論文では，できるかぎりレポートの主題は似ているが，内容が異なるレポートを推薦する手法を提案する．同じ主題の 2 つのレポートの内容が異なるほど，それらのレポートライティングにおける多様なスキルが異なる確率が高まると考えられる．提案手法では，他者のレポートを学習者に推薦し，自分のレポートと比較することにより，レポートの内容を深く推敲する機会を多く作るだけでなく，他者のレポートライティングにおける多様なスキルを学ぶことができると考える．

第 2 章では，本推薦システムで用いる学習者のレポートデータを蓄積している LMS（Learning Management System）“Samurai”に，レポート推薦システムの関連研究を紹介する．推薦システムの関連研究を，レポートライティング支援システムと教育分野における推薦システムに大別し紹介する．多くのレポートライティング支援システムは，「導入，背景，目的，方法，結論」と

いった形式的な構成を解析し，学習者の論文構成を可視化や指摘するシステムが多い．教育分野における推薦システムは，機械学習手法や時系列モデル，オントロジー手法を用い，学習者の学力や興味に応じたコンテンツを推薦するシステムである．このような従来の推薦システムは，いずれも学習者データと類似性が高いコンテンツや人，メッセージを推薦しており，レポート推薦に用いると類似したものばかりが推薦されてしまい，学習者のレポートとの差異が少なく，学習効果が少ないと考えられる．そのため，レポートライティングにおける推薦手法を第 3 章において提案する．

第 3 章では，LDA を用いたレポート推薦システムを提案する．使用したレポートは，LMS”Samurai”に蓄積されているレポートデータを用いた．本提案システムは，レポートライティングにおける「他者からの学び」を支援することを目的している．そのため，従来のレポートライティング支援システムのような学習者のレポートの形式的な構成を解析し，学習者が着目すべき箇所を指摘する手法ではなく，学習者に他者のレポートそのものを推薦する．その際，どのようなレポートを推薦することで，学習者に有用であるかが問題となる．本章では，技術的には，Latent Dirichlet Allocation（LDA）を用いることにより，できるかぎり主題は似ているが内容（用いられる単語分布）が異なるレポートを推薦する手法を提案する．これにより，主題は同じでもレポートライティングにおける多様なスキルを持つレポートが推薦できると期待できる．ただし，ここでいう「構成」とは「導入，背景，目的，方法，結論」などといった形式的な構成ではなく，レポートの主張点の論理構成や文章の流れを意味する．また，実際の理工系大学生を対象に評価実験を行い，本提案の有効性を示した．

第 4 章では，LDA を用いる際に，予め決定しておく必要があるトピック数の決定手法について述べる．第 3 章において，これまでトピック数をデータから推定する手法が確立されていなかったので，第 3 章ではトピック数を決めて用いている．しかし，データが大量になった場合や新たにデータを追加する際に人手によりレポートを分類し，トピック数を決める必要があり，シス

テムを利用する上で現実的ではない．また，人手による分類に即したトピック数が，モデルの学習・推定精度を高くする保証はない．そこで，本章では，トピック数を変え，LDA の周辺尤度を計算し，周辺尤度の値が最も高くなるときのトピック数をモデルの真のトピック数として採用する．周辺尤度を計算する際，LDA のハイパーパラメータが結果に大きく影響することをシミュレーションにより示した．結果として，LDA のハイパーパラメータを 1 としたときに，LDA のトピック数を推定できることをシミュレーションにより示した．この結果を本推薦システムに組み込むことで，その有効性を示した．

最後に第 5 章では，本研究で得られた主な研究成果を統括し，本論文をまとめるとともに本研究の課題について述べる．

第 2 章

関連研究

本章では，本推薦システムで用いる学習者のレポートデータを蓄積している LMS“Samurai”，レポート推薦システムの関連研究を紹介する．推薦システムの関連研究を，レポートライティング支援システムと教育分野における推薦システムに大別し紹介する．多くのレポートライティング支援システムは，「導入，背景，目的，方法，結論」といった形式的な構成を解析し，学習者の論文構成の可視化や修正すべき箇所を指摘するシステムである．教育分野における推薦システムは，機械学習手法や時系列モデル，オントロジー手法を用い，学習者の学力や興味に応じたコンテンツを推薦するシステムである．このような従来の推薦システムは，いずれも学習者データと類似性が高いコンテンツや人，メッセージを推薦している．

2.1 LMS“samurai”

本論文では，植野ら [1–5] が長年開発してきた LMS（Learning Management System）“Samurai” に蓄積された学習者データを用いる．LMS“Samurai”では，学習者がメニュー画面より，学習コンテンツを選ぶことで学習を進める．各コンテンツは，教師映像と説明用テキスト画面，説明用ビデオ映像，演習用テスト，課題により構成され，掲示板システムにより，課題

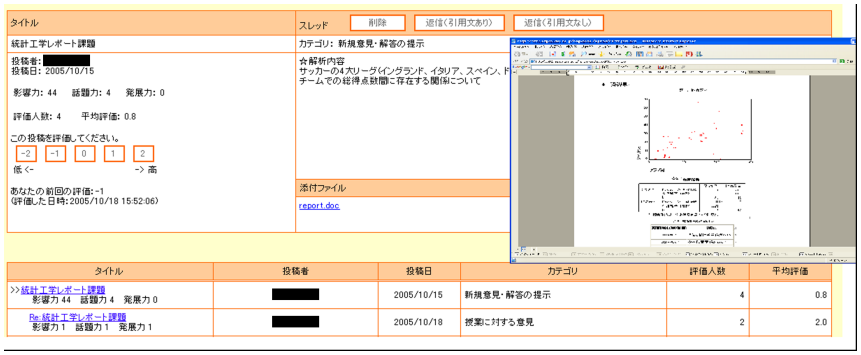

図 2.1.1 LMS“Samurai”内の掲示板

提出，学習者間の議論，ピアレビュー等ができる．学習者が掲示板に投稿したレポートに対し，他の学習者による評価，学習者間での議論が行われる．学習者が提出した課題レポートやテストの成績，回答時間，議論などの履歴は，学習履歴データベースへ自動的に格納される．図 2.1.1 は，“Samurai”内の掲示板を示す．本論文では，これら過去に蓄積された他者のレポートを，初心者のレポートライティングの学習に利用する．

2.2 レポートライティング支援システム

これまで，レポートライティングを支援するシステムが多数開発されている．例えば，O’Rourke and Calvo [6] は，段落間の関係性を可視化するシステムを開発している．西村ら [7]，甲斐ら [8] は，文章の表現から論文構成を解析するシステムを開発している．岩田ら [9]，山崎ら [10] は論文構成の規範と利用者の論文構成を比較することで，利用者の論文構成を指摘するシステムを開発している．Toulmin モデルに論証を当てはめ可視化するシステムが開発されている [11, 12]．宇都と植野は，確率的アプローチを用いて論文構成の構築過程を支援するシステム [13]，Toulmin モデルのベイジアンネットワーク表現を用いた論証構築支援システムを開発している [14]．

しかし，これらは「導入，背景，目的，方法，結論」などの論文の文章構

造の構築を形式的に支援するものである．本提案では，他者のレポートを学習者に推薦し，自分のレポートと比較することにより，レポートの内容を深く推敲する機会を多く作るだけでなく，他者のレポートライティングの方法を学ぶことができると考える．この場合，どのように学習者にレポートを推薦するかが問題である．

2.3 教育分野における推薦システム

これまでに教育分野では, 多くの推薦システムが開発されている [15]．具体的には，機械学習手法や時系列モデル，オントロジー手法を用い，学習者の学力や興味に応じたコンテンツを推薦するシステムである．例えば，論文を推薦するシステムとしては，[16–18] などがある．一例としては，Bollacker et al. [16] の閲覧論文と TFIDF（Term Frequency Inverse Document Frequency）[19] を用いた類似度の高い論文を推薦するシステムがある．TFIDF は，文書中に含まれる特徴的な単語に重みづけをする手法である．

学習コンテンツを推薦するシステムとしては，[20–25] などがある．例えば，Khribi et al. [23] は，学習者の学習履歴を基に学習教材を推薦するシステムを提案している．Yang ら [25] は，ビデオ教材の映像に付与されているテキスト情報と学習履歴の TFIDF を用いた類似度が高いビデオ教材を推薦するシステムを提案している．

参考文献や学習コンテンツではなく，学習の過程そのものを推薦するシステムも近年提案されている [4, 26]．例えば，Huang et al. [26] は，マルコフ連鎖モデルを用いて学習プロセスをモデル化し，推薦を行うシステムを開発している．植野と宇都 [4] は，学習履歴そのものである e ポートフォリオを推薦するシステムを提案している．このシステムは，単純に評価の高い学習者の e ポートフォリオを推薦するのではなく，対象学習者と推薦する e ポートフォリオとの類似性を考慮した推薦を行うことで，学習効果を高めている．

このような従来の推薦システムは，いずれも学習者データと類似性が高い

コンテンツや人，メッセージを推薦している．しかし，このような従来手法をレポート推薦に適用する場合，内容・表現が類似のレポートばかりが推薦されてしまい，効果的な学習が期待できない．

2.4 むすび

本章では，LMS“Samurai”，レポートライティング支援システム，教育分野における推薦システムについて紹介した．多くのレポートライティング支援システムは，「導入，背景，目的，方法，結論」といった形式的な構成を解析し，学習者の論文構成を可視化や指摘するシステムである．教育分野における推薦システムは，いずれも学習者データと類似性が高いコンテンツや人，メッセージを推薦している．

本論文の主なアイデアは，できるかぎり主題は似ているが内容（用いられる単語分布）が異なるレポートを推薦する手法である．これにより，主題は同じでも様々な構成や表現，オリジナリティのレポートが推薦できると期待できる．ただし，ここでいう「構成」とは「導入，背景，目的，方法，結論」などといった形式的な構成ではなく，レポートの主張点の論理構成や文章の流れを意味する．技術的には，文書のトピック（潜在的な意味）を推定できる Latent Dirichlet Allocation(LDA) [27] を用いて，学習者と他者のレポート間のトピック分布距離を計算することで類似の主題を持つレポートを同定する．LDA を用いて推定されるトピックは，意味が同じで異なる単語も同一のトピックとして推定できる．すなわち，LDA はトピック分布と表層的な単語分布を分離して扱うことができることが特徴ともいえる．本論文では，同じ主題であれば，単語分布がレポートの内容を反映していると仮定する．同じ主題の 2 つのレポートの単語分布が異なるほど，それらの「構成」,「表現」,「オリジナリティ」が異なる確率が高まる．そのために，トピック分布が類似で異なる単語分布のレポートを推薦すれば，多様な「構成」,「表現」,「オリジナリティ」を持つレポートを推薦できると考えられる．学習者のレポートライティングにお

ける「構成」,「表現」,「オリジナリティ」についての能力を向上させると予想できる．そして，他者からの学びは，単一の他者のみからよりも多様な他者からの学びの方が効果的であることが知られており [28]，提案手法により，より効率的な学習ができると期待できる．第 3 章では，その具体的な方法について記述する．

第 3 章

LDA を用いたレポート推薦システム

3.1　はじめに

近年，高等教育におけるライティング教育の重要性が指摘されている [29]. しかし，初心者には，独力でレポートを書き上げることは難しい．本論文では，徒弟的アプローチ [30] に基づき，過去の優秀なレポートを適応的に推薦することにより，初心者のレポートライティングを支援する手法を提案する．

近年，学習理論の主流は，Vygotsky に代表される社会的構成主義 [30] に移行しつつある． Vygotsky は，人の知識構築は単なる知識の伝達ではなく，

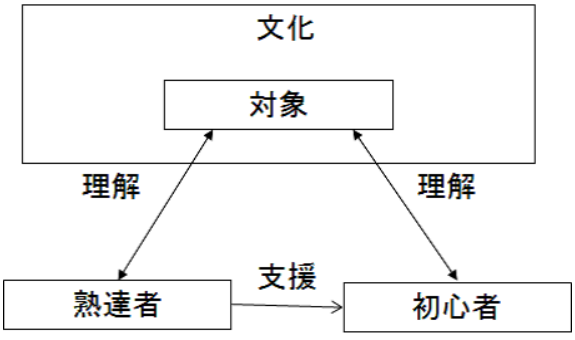

図 3.1.1　Vygotsky の学習モデル

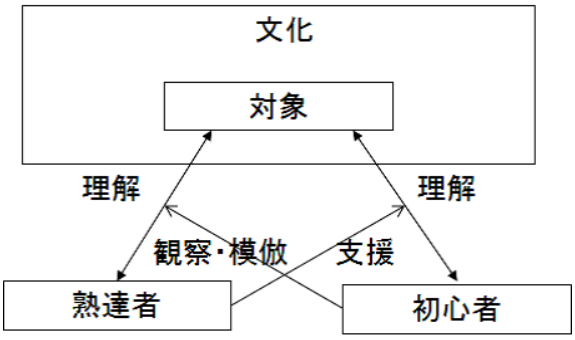

図 3.1.2　植野の Vygotsky モデルの解釈

図 3.1.1 のような，対象の理解の仕方への支援としてモデル化している．初心者は，熟達者に問題解決や対象理解を支援してもらうことにより，最初は表層的ではあるが，徐々に, 単なる知識のみでなく，理解の仕方，注意・焦点化，内省，態度，動機，情熱などの対象に関する高次の心的スキルを獲得できると主張している．このモデルに従えば，教師は学習対象の面白さや情熱，見方や価値観，倫理，その背景，文化を伴って支援するので，教師の対象の見方そのものを獲得できる．また，初心者は熟達者から一方的に支援されるのではなく，意識的に他者から学ぼうとしており，観察や模倣，他者との比較などが行われる．学習者の発達に伴い，熟達者の支援がなくても自律的に他者からの学びが行われると考えられる．植野 [31] は，図 3.1.2 において，初心者は熟達者から支援されることが主であるが，徐々に発達して学習者自身からの観察・模倣といった自律的な他者からの学びができるように変化すると述べている．そして，この変化が発達の本質であると述べている．本論文では，このモデルに従い，レポートライティングにおける「他者からの学び」を支援するシステムを提案する．具体的には，過去の熟達者のレポートを学習者に適応的に推薦し，レポートライティングにおける「他者からの学び」を支援する．

第 2 章において，従来のレポートライティング支援システムを紹介した．しかし，これらは「導入，背景，目的，方法，結論」などの論文の文章構造の構築を形式的に支援するものである．本提案では，他者のレポートを学習者に

推薦し，自分のレポートと比較することにより，レポートの内容を深く推敲する機会を多く作るだけでなく，他者のレポートライティングの方法を学ぶことができると考える．この場合，どのように学習者にレポートを推薦するかが問題である．第 2 章において，教育分野における推薦システムの関連研究を紹介した．教育分野における従来の推薦システムは，いずれも学習データと類似性が高いコンテンツを推薦している．しかし，このような従来手法をレポート推薦に適用する場合，内容・表現が類似のレポートばかりが推薦されてしまい，効果的な学習が期待できない．レポートライティングにおける「他者からの学び」を支援するためには，できるだけ学習者のレポートの内容と差異があることが望ましい．しかし，レポートの主題はできるだけ似ているものであることが望ましい．

そこで本論文では，できるかぎり主題は似ているが内容（用いられる単語分布）が異なるレポートを推薦する手法を提案する．これにより，主題は同じでも様々な構成や表現，オリジナリティのレポートが推薦できると期待できる．ただし，ここでいう「構成」とは「導入，背景，目的，方法，結論」などといった形式的な構成ではなく，レポートの主張点の論理構成や文章の流れを意味する．技術的には，文書のトピック（潜在的な意味）を推定できる Latent Dirichlet Allocation(LDA) [27] を用いて，学習者と他者のレポート間のトピック分布距離を計算することで類似の主題を持つレポートを同定する．LDA を用いて推定されるトピックは，意味が同じで異なる単語も同一のトピックとして推定できる．すなわち，LDA はトピック分布と表層的な単語分布を分離して扱うことができることが特徴ともいえる．本論文では，同じ主題であれば，単語分布がレポートの内容を反映していると仮定する．同じ主題の 2 つのレポートの単語分布が異なるほど，それらの「構成」,「表現」,「オリジナリティ」が異なる確率が高まる．そのために，トピック分布が類似で異なる単語分布のレポートを推薦すれば，多様な「構成」,「表現」,「オリジナリティ」を持つレポートを推薦できると考えられる．学習者のレポートライティングにおける「構成」,「表現」,「オリジナリティ」についての能力を向上させると予想できる．そして，

他者からの学びは，単一の他者のみからよりも多様な他者からの学びの方が効果的であることが知られており [28]，提案手法により，より効率的な学習ができると期待できる．

実際の理工系大学生を対象に評価実験を行い，本提案の有効性を示した．レポートデータは，LMS（Learning Management System）“Samurai” [1–5] に蓄積された学習者データを用いる．

3.2 Latent Dirichlet Allocation（LDA）

本節では，レポートの主題を推定するために用いるトピックモデルについて述べる．トピックモデルとは，文書中の単語は文書の潜在的な意味 (トピック) に依存して出現すると仮定し，文書中に出現する単語の頻度からそのトピックを推定する手法である．トピックモデルの代表例として Latent Semantic Analysis（LSA）[32]，Probabilistic Latent Semantic Indexing（PLASI）[33]，Latent Dirichlet Allocation（LDA） [27] がある．LDA は，LSA と PLSI よりもトピックを高精度に推定することが可能であり，計算効率も良いことが知られている [27]．そのため，本論文では LDA を採用する．

トピックモデルを教育分野に応用した研究としては，椿本らの [34, 35] がある．これらの研究は，評価者のレポート採点時の評価基準の曖昧さを軽減するシステムを提案している．しかし，これは本研究の目的とは異なる．

LDA は文書が生成される過程を確率的に表現したモデルである．一つの文書が複数の潜在的意味 (トピック) を持つと仮定する．各文書は文書内の含まれるトピックの割合を示すトピック分布 θ を持つ．θ に従い文書内のトピック z が選ばれる．トピックが選ばれると，トピックに対応する単語の分布 ϕ に従い単語が生成される．θ は各文書ごとにディリクレ分布から生成され，ディリクレ分布のパラメータ α をハイパーパラメータと呼ぶ．ϕ は各トピックごとにディリクレ分布から生成され，ハイパーパラメータは β である．

LDA のグラフィカルモデルは図 3.2.1 のように表される．図 3.2.1 において，W は観測される文書内の単語を示す．また，K はトピック数，D は文書

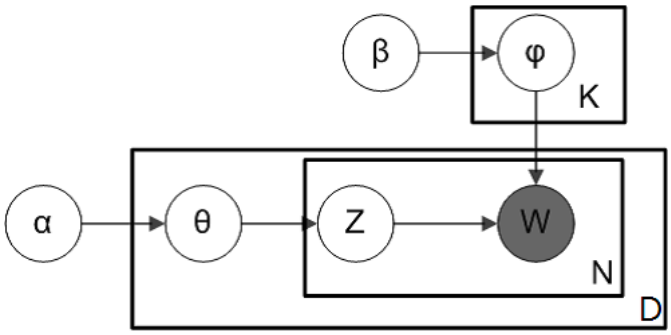

図 3.2.1　LDA のグラフィカルモデル

数，N を文書内の単語数，Z はトピック，ϕ_k はトピック k が持つ語彙配分，θ_d は文書 d が持つトピック配分を表す．α，β は，ディリクレ事前分布のパラメータであり，ハイパーパラメータと呼ぶ．トピックモデルにおける文書集合 W とトピック集合 $Z = \{\{z_{dn}\}_{n=1}^{N_d}\}_{d=1}^{D}$ の事後分布は式（3.1），式（3.2），式（3.3）で表わされる．

$$P(W, Z \mid \alpha, \beta) = P(Z \mid \alpha) P(W, Z \mid \beta), \tag{3.1}$$

$$P(Z \mid \alpha) = \left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha)^K} \right)^D \prod_{d=1}^{D} \frac{\prod_{k=1}^{K} \Gamma(N_{kd} + \alpha_k)}{\Gamma(N_d + \sum_{k=1}^{K} \alpha_k)}, \tag{3.2}$$

$$P(W \mid Z, \beta) = \left(\frac{\Gamma(\sum_{v=1}^{V} \beta_v)}{\prod_{v=1}^{V} \Gamma(\beta_v)} \right)^K \prod_{k=1}^{K} \frac{\prod_{v=1}^{V} \Gamma(N_{kv} + \beta_v)}{\Gamma(N_k + \sum_{v=1}^{V} \beta_v)}. \tag{3.3}$$

ここで $\Gamma(\cdot)$ はガンマ関数を表す．V は語彙数，$N_k d$ は文書 d に含まれるトピック k の数を示す．$N_d = \sum_{k=1}^{K} N_{kd}$ を満たす．N_{kv} はトピック k に割り当てられた語彙 v の数を示す．$N_k = \sum_{v=1}^{V} N_{kv}$ を満たす．

文書 d におけるトピック分布を θ_d，トピック k のときの単語配分を ϕ_k と表すとき，それぞれ下式により推定できる．

$$\hat{\theta}_d = \frac{N_{kd} + \alpha}{N_d + K\alpha}, \tag{3.4}$$

$$\hat{\phi}_k = \frac{N_{kv} + \beta}{N_k + V\beta}. \tag{3.5}$$

式（3.4），式（3.5）は，文書の単語の頻度情報を入力として，崩壊型ギブスサンプリングを用い推定することができる [36].

ハイパーパラメータ α，β は，不動点反復法 [37] を用いて周辺尤度を最大

化することによりデータから推定できる．α，β は下式により更新される．

$$\alpha^{new} \longleftarrow \alpha \frac{\sum_d D \sum_k K(\Psi(N_{kd}+\alpha) - \Psi(\alpha))}{K \sum_d D(\Psi(N_d + K\alpha) - \Psi(K\alpha))} \tag{3.6}$$

$$\beta^{new} \longleftarrow \beta \frac{\sum_k K \sum_v V(\Psi(N_{kv}+\beta) - \Psi(\beta))}{V \sum_k K(\Psi(N_k + V\beta) - \Psi(V\beta))} \tag{3.7}$$

ここで，$\Psi(x)$ はディガンマ関数を示す．

3.3 LDA モデルの学習手法

LDA における代表的な学習手法である，変分ベイズ法 [27]，崩壊型ギブスサンプリング [36] を紹介する．

3.3.1 変分ベイズ法

LDA における変分ベイズ法 (Variational Bayes Inference) [27] について述べる．

LDA の学習は，文書データ W が与えられた時の潜在変数 Z の事後分布を計算することが目的である．しかし直接計算することは困難である．変分ベイズ法はこの問題を解決するために，確率変数 z, θ, ϕ が互いに独立であると仮定している．この仮定の下で $q(z,\theta,\phi) = \prod_z q(z) \prod_d q(\theta_d) \prod_k q(\phi_k)$ と $p(z, \theta_d, \phi_k \mid W)$ とのカルバックライブラーダイバージェンスを最小化するように $q(z,\theta,\phi)$ を求める手法である．しかし，実際には z, θ, ϕ は互いに独立ではなく依存関係にある．q を直接求めることが難しいため，変分ベイズ法を用いて近似し，その下界を最大化することを考える．LDA における変分ベイズ法によるパラメータ推定の裏付けとなる Jensen の不等式は，文書 d_i が生成される確率を $P(d_i \mid \alpha, \beta)$ とし，文書 d_i の各単語へのトピックの割り当てを z_i

として，以下のように表せる．

$$\begin{aligned}\log P(d_i \mid \alpha, \beta) &= \log \int \sum_{z_i} P(\theta, z_i, d_i \mid \alpha, \beta) d\theta \\ &= \log \int \sum_{z_i} \frac{P(\theta, z_i, d_i \mid \alpha, \beta) Q(\theta, z_i \mid \gamma, \phi)}{Q(\theta, z_i \mid \gamma, \phi)} d\theta \\ &\geq \int \sum_{z_i} Q(\theta, z_i \mid \gamma, \phi) \log P(\theta, z_i, d_i \mid \alpha, \beta) d\theta \\ &- \int \sum_{z_i} Q(\theta, z_i \mid \gamma, \phi) \log Q(\theta, z_i \mid \gamma, \phi) d\theta \end{aligned} \tag{3.8}$$

ここで $Q(\theta, z_i \mid \gamma, \phi)$ は，$P(\theta, z_i, d_i \mid \alpha, \beta)$ を近似するために導入された確率分布であり，互いに独立な項の積で表されていると仮定する．つまり，

$$Q(\theta, z_i \mid \gamma, \phi) = Q(\theta \mid \gamma) \prod_{l=1}^{n_i} Q(z_{il} \mid \phi_l) \tag{3.9}$$

と表されると仮定する．ここで，n_i は文書 d_i の長さ，z_{il} は文書 d_i における第 l 番目の単語のトピックを表し，ϕ_l は文書 d_i における第 l 番目の単語のトピックを定める多項分布のパラメータである．つまり $\phi_l, l = 1, ..., n_i$ は，トピックの総数を K として，K 個のパラメータ $\phi_{l1}, ..., \phi_{lK}, s.t. \sum_k \phi_{lk} = 1$ の集まりである．LDA 文書モデルにおいて，各文書 d_i における各単語のトピック $z_i = \{z_{il}, ..., z_{in_i}\}$ を定める多項分布 $P(z_i \mid \theta)$ は，トピックの事前分布 $P(\theta \mid \alpha)$ に依存している．このため異なる文書におけるトピックの出現確率の分布 $P(z_i \mid \theta), i = 1, ..$ を別々に扱うことはできない．しかし変分ベイズ法では，$Q(\theta \mid \gamma)$ と $Q(z_{il} \mid \phi_l)$ は互いに独立と仮定する．これはパラメータ推定が各文書について別々に行われることを意味する．よって，γ, ϕ も各文書 d_i ごとに別々に推定される．これは変分ベイズ法を用いることの利点である．

特定の文書 d_i に対して $\log P(d_i \mid \alpha, \beta)$ を最大化したいのであるが，直接最大化することが困難である．そこで変分ベイズ法を用いることで代わりに以下の不等式の右辺に与えられている下界を最大化することでパラメータ推定を

行う．

$$\begin{aligned}\log P(d_i \mid \alpha, \beta) \geq \log\Gamma(\sum_k \alpha_k) - \sum_k \log\Gamma(\alpha_k) \\ + \sum_{k'}(\alpha_{k'} - 1)(\Psi(\gamma'_k) - \Psi(\sum_k \gamma_k)) + \sum_l \sum_{k'} \phi_{lk'}(\Psi(\gamma'_k) - \Psi(\sum_k \gamma_k)) \\ + \sum_l \sum_j \delta_{lj} \sum_k \phi_{lk} \log \beta_{kj} - \log\Gamma(\sum_k \gamma_k) + \sum_k \log\Gamma(\gamma_k) \\ - \sum_{k'}(\gamma'_k - 1)(\Psi(\gamma'_k) - \Psi(\sum_k \gamma_k)) - \sum l \sum_k \phi_{lk} \log \phi_{lk}\end{aligned}$$

これを最大化するような ϕ, γ を求めればよい．ϕ_{lk} は，文書 d_i における第 l 番目の単語のトピックが k となる確率を表すために導入された，変分パラメータである．γ_k は，変分法を用いる際に導入したトピックのディリクレ事前分布のパラメータである．ϕ_{lk}, γ_k で偏微分し，それぞれの式が 0 に等しいとすると，以下のように計算できる．

$$\phi_{lk} = \beta_{kjl} exp(\Psi(\gamma_k) - \Psi(\sum_k{}')\gamma_k = \alpha_k + \sum_{l=1}^{n_i} \phi_{lk} \tag{3.10}$$

また α, β は以下の更新式により求められる．

$$\beta_{kj} \propto \sum_{d=1}^{M} \sum_{n=1}^{N_d} \phi_{dni} w_{dn}^j \tag{3.11}$$

$$\begin{aligned}\alpha_k = \hat{\alpha_k} + \left(\frac{\Psi(\sum_k \hat{\alpha_k})}{\Psi_1(\hat{\alpha_k})} - \frac{\Psi(\hat{\alpha_k})}{\Psi_1(\hat{\alpha_k})} + \frac{\sum_i (\Psi(\gamma_{ik}) - \Psi(\sum_k \gamma_{ik}))}{N\Psi_1(\hat{\alpha'_k})} \right) \\ + \left(\frac{\Psi_1(\sum_k \hat{\alpha_k})}{\Psi_1(\hat{\alpha_k})} - \sum_{k'} \frac{\Psi_1(\sum_k \hat{\alpha_k})}{\Psi_1(\hat{\alpha_k})} \right)^{-1} \qquad (3.12) \\ \times \sum_{k'} \left(\frac{\Psi(\sum_k \hat{\alpha_k})}{\Psi_1(\hat{\alpha_k})} - \frac{\Psi(\hat{\alpha_k})}{\Psi_1(\hat{\alpha_k})} + \frac{\sum_i (\Psi(\gamma_{ik}) - \Psi(\sum_k \gamma_{ik}))}{N\Psi_1(\hat{\alpha'_k})} \right) \qquad (3.13)\end{aligned}$$

$\Psi_1(x)$ はディガンマ関数 $\Psi(x)$ の微分であり，トリガンマ関数である．

3.3.2 崩壊型ギブスサンプリング

LDA における崩壊型ギブスサンプリング (Collapsed Gibbs Sampling) [36] について述べる．LDA に基づく予測には，データが与えられた時の $p(Z \mid W)$ を推定すればよい．$p(Z \mid W)$ に従うサンプルが得られれば，文書 d におけるトピック k が生成される確率の推定量である $\hat{\theta}_{kd}$ や，トピック k から語彙 v が生成される確率の推定量である $\hat{\phi}_{kv}$ が計算できる．そこで $p(Z \mid W)$ に従うサンプルを得ることが目的になる．崩壊型ギブスサンプリングでは，確率変数 z の成分 z_i に関する条件付き分布 $p(z_i \mid z_{\setminus i}, W)$(あるいはそれに比例する関数 $q_i(z_i)$) を使って，マルコフ系列を作り，それの部分列をサンプルとして使う．ギブスサンプリングはマルコフ連鎖モンテカルロ法の一種である．条件付き確率そのものではなく，それに比例する関数 $q_i(z_i)$ が与えられればサンプルを作ることが出来る．すなわち $p(z_i = j \mid z_{\setminus i}, w)$ において異なる j の間での相対的な大小関係が分かればよい．$z_{\setminus i}$ は，z から i 番目の z_i を除くという意味で用いた．

トピック集合 Z は，文書集合 W を入力とし，崩壊型ギブスサンプリングを用いることで効率的に推定できる．文書 d の n 番目を生成する単語のトピック z_j，$j = (d, n)$ のサンプリング確率は下式により計算できる．

$$P(z_j = k \mid Z_{\setminus j}, W) \propto \frac{N_{kd\setminus j} + \alpha_k}{N_{d\setminus j} + \sum_{k=1}^{K} \alpha_k} \cdot \frac{N_{kv\setminus j} + \beta_v}{N_{k\setminus j} + \sum_{v-1}^{V} \beta_v} \tag{3.14}$$

ここで N_{kd} は文書 d におけるトピック k が割り当てられた単語数，N_{kw} はトピック k における単語 w の出現回数を表す．N_k はトピックコーパス z においてトピック k が表れた回数を示し $N_k = \sum_{v=1}^{V} N_{kv}$ である．N_d は文書 d に含まれる単語の数を示し，$N_d = \sum_{k=1}^{K} N_{kd}$ である．$N_{d\setminus j}$ は文書 d の n 番目の単語を除いたときの単語の数を表す．式 (3.14) は，文書 d でのトピック k の割合と，トピック k での語彙 v の割合の積で表されている．崩壊型ギブスサンプリングの計算量は $\mathcal{O}(NK)$ である．ただし，N は全文書の全単語数を示し，K はトピック数を示す．変分ベイズ法よりも崩壊型ギブスサンプリングの方が

実装が容易であり，計算速度が速く，精度も高いことが知られている [38]．これらの利点から，本論文では LDA の学習手法に崩壊型ギブスサンプリングを採用する．

3.4 LDA によるデータ分析

本節では，第 2 章で紹介した LMS“Samurai” に蓄積された学習者のレポートデータに対して，Latent Dirichlet Allocation(LDA) [27] を用いて分析する．まず，類似度算出手法について紹介する．

3.4.1 類似度算出手法

レポート推薦のために，文書間の主題の類似性及び表面的な出現単語の類似性を定義する．LDA の技術的な利点の一つは，文書の主題を反映するトピックの確率分布と，文書で用いられた単語の確率分布を別々に扱うことができる点である．本論文では，この性質を用いて文書間の主題の非類似度（距離）と出現単語の非類似度（距離）を，トピック分布と単語分布それぞれの Jsensen-Shannon ダイバージェンスにより定義する．また，比較のため，文書間の内容の類似度を評価する従来手法である TFIDF を用いるコサイン類似度についても本節で紹介する．

Jensen-Shannon ダイバージェンス

確率分布間の非類似度（距離）を示す指標として，Jensen-Shanon ダイバージェンスを紹介する．この指標は，2 つの確率分布が一致するとき最小値 0 をとり，異なれば異なるほど大きな正の値を返す擬似距離である．

Kullback-Leibler ダイバージェンスを KLD で表わすとき，文書 d_i, d_j 間のトピック分布の Jensen-Shannon ダイバージェンス（$\mathrm{T_{JSD}}$）は，次式で表わされる．

$$\mathrm{T_{JSD}(d_i, d_j)} = \frac{1}{2}\mathrm{KLD}(\theta_{\mathrm{d_i}} \parallel \mathrm{m}) + \frac{1}{2}\mathrm{KLD}(\theta_{\mathrm{d_j}} \parallel \mathrm{m}) \tag{3.15}$$

ここで，$\mathrm{KLD}(\theta_{\mathrm{d_i}} \parallel \mathrm{m}) = \sum_{\mathrm{k}} \theta_{\mathrm{d_i,k}} \ln \frac{\theta_{\mathrm{d_i,k}}}{\mathrm{m}}$，文書 d_i のトピック分布を $\theta_{d_i} = [\theta_{d_i,k=1}, \ldots, \theta_{d_i,k=K}]$ とし，$m = \frac{1}{2}(\theta_{d_i} + \theta_{d_j})$ とする．これにより，2 文書間のトピック分布の距離が求められ，2 文書が同一のトピック分布を持つ場合には 0 となる．これを用いることで，対象レポートと同一主題のレポートを探し出すことができる．

文書 d_i, d_j 間の単語分布の Jensen-Shannon ダイバージェンス（$\mathrm{W_{JSD}}$）は，次式で表わされる．

$$\mathrm{W_{JSD}(d_i, d_j)} = \frac{1}{2}\left(\mathrm{KLD(wd_{d_i} \parallel l)} + \mathrm{KLD(wd_{d_j} \parallel l)}\right) \tag{3.16}$$

ここで，$\mathrm{KLD(wd_{d_i} \parallel l)} = \sum_{\mathrm{v}} \mathrm{wd_{d_i,v}} \ln \frac{\mathrm{wd_{d_i,v}}}{\mathrm{l}}$，文書 d_i の単語分布を $wd_{d_i} = [N_{d_i,v=1}/N_{d_i}, \ldots, N_{d_i,v=V}/N_{d_i}]$，$N_{d_i,v}$ は，文書 d_i における単語 v の出現頻度，N_{d_i} は文書 d_i 内の単語総数を示す．また $l = \frac{1}{2}(wd_{d_i} + wd_{d_j})$ とする．これは 2 文書間で用いられている単語分布の距離を評価する指標であり，同一の単語分布を持っている場合には 0 の値になる．2 文章間の表層的な単語出現の仕方による違いを示し，対象レポートとなるべく異なる表現方法のレポートを探し出すのに用いられる．

コサイン類似度

TFIDF（Term Frequency Inverse Document Frequency）による文書間の類似度にはコサイン類似度を用いる．TFIDF は，文書中に含まれる特徴的な単語に重みづけをする手法である．文書 d における単語 v の TFIDF 値は，以下のように定義される．

$$\mathrm{TFIDF(v, d)} = \frac{\mathrm{N_{dv}}}{\mathrm{N_d}} \cdot (\ln \frac{\mathrm{D}}{\mathrm{df(v)}} + 1) \tag{3.17}$$

N_{dv} は文書 d における単語 v の頻度，N_d は文書 d における単語数，D は文書数，$df(v)$ は，単語 v が出現する文書数を示す．

項目 d_i, d_j 間のコサイン類似度 (CosSim) は，以下のように表わせる．

$$\mathrm{CosSim(d_i, d_j)} = \frac{\mathrm{TFIDF_{d_i}} \cdot \mathrm{TFIDF_{d_j}}}{\parallel \mathrm{TFIDF_{d_i}} \parallel \parallel \mathrm{TFIDF_{d_j}} \parallel} \tag{3.18}$$

ここで，$\mathrm{TFIDF_{d_i}}$ は文書 d_i の TFIDF 値のベクトルを示し，$\mathrm{TFIDF_{d_i}} = [\mathrm{TFIDF_{d_i,v=1}}, \ldots \mathrm{TFIDF_{d_i,v=V}}]$ と表す．この指標は，0 から 1 までの値を示

し，類似度が高いと 1 に近づく．

3.4.2 LDA による分析

データ

前述のように”Samurai”内には，実際の講義の課題として提出されたレポートが蓄積されている．ここでは，理工系大学の修士課程の講義「知識創産システム論」における 90 のレポートについて LDA で分析した．全てのレポートの語彙数は 5492，単語数は 16796 であった．講義でのレポート課題は「企業における従来の知識創産手法とその問題点について述べよ」として提示された．

LDA に代表されるトピックモデルは，文書中の単語の語彙数と頻度情報からトピックを推定する．そのため，LDA に文書データを入力する前処理として，分かち書きにより単語区切りに分割する必要がある．本研究では，形態素解析器 MeCab [39] を用いて，各レポートに対して分かち書きを行った．また，ストップワードと呼ばれる言語的に意味のない語を除外した．例えば，「そして」，「つまり」などの接続詞や，「の」，「に」，「と」などの助詞を指す．

3.4.3 レポートデータのトピック数の推定

データから LDA のトピック数を決定するために，一般的に，モデル選択基準であるベイズ情報量基準（BIC），赤池情報量規準 (AIC)，周辺尤度を用いる．BIC や周辺尤度は，データ数に対して漸近一致性を持つが，LDA での推論を最適化できない場合が多い．そこで，本論文では分類精度の尺度である F 値を用いる．具体的には，各トピック数毎に LDA により推定されたトピック分布を用い，レポート間のトピック分布の類似度を式（3.15）から算出する．k-means 法 [40] によりレポートを分類し，人の手による分類との一致精度（F 値）を求めた．F 値は $F = 2rp/(r + p)$ で表わされ，r は再現率（正解データのうち，正解であると認識された割合）を示し，p は適合率（正解であると認識

したデータのうち，正解であるデータの割合）を示す．正解データはレポート課題の専門家にレポートを分類してもらい，作成した．図 3.4.1 は，各トピック数毎に算出した F 値の最大値を示す．トピック数 $K = 4$ のときの F 値が最大値を示したため，トピック数を 4 とした．表 3.1 は，トピック数が 4 のときの各レポートの主題毎の再現率，適合率を示す．主題のひとつであるナレッジマネジメントのレポート数が他の主題に比べて少ないため，トピック分布の推定精度が低くなり，他の主題と比べての適合率が低くなったと考えられる．その他のレポートの再現率・適合率の値は高い値を示しているため，正解データとトピック分布による分類の差は小さいと考えられる．

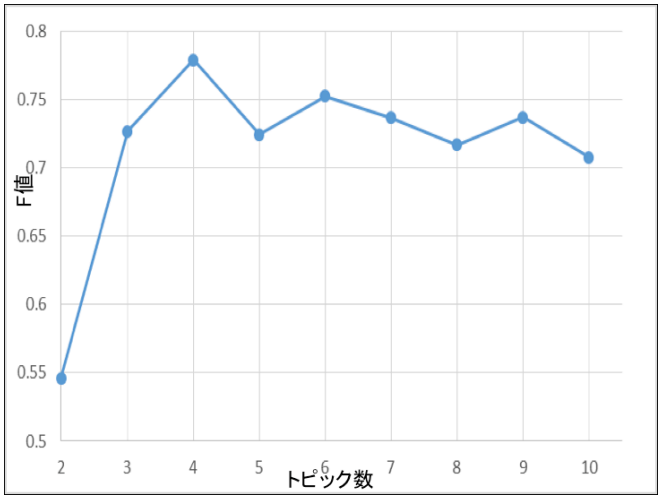

図 3.4.1　各トピック数での F 値の最大値

表 3.1　トピック数 4 のときトピック分布による分類結果 (再現率・適合率)

主題（レポート数）	再現率	適合率
科学的管理論 (24)	1	0.92
産業革命（27）	0.85	0.96
ナレッジマネジメント（7）	0.75	0.46
リエンジニアリング（32）	0.77	0.89

表 3.2　推定された各トピックの単語

トピック	単語 (出現確率)
トピック 1 科学的管理論	管理 (0.0340), 労働 (0.0258), 科学 (0.0216), 作業 (0.0176), テーラー (0.0176), 実践 (0.0101), 生産 (0.0076), 仕事 (0.0076), 経営 (0.0069), システム (0.0069)
トピック 2 産業革命	技術 (0.0167), 企業 (0.0161), 産業 (0.0139), 社会 (0.0116), 革命 (0.0115), 情報 (0.01074), ベンチャー (0.0104), 日本 (0.0093), 精神 (0.0087), 知識 (0.0085)
トピック 3 ナレッジマネジメント	知識 (0.0133), 看護 (0.0083), ます (0.0083), 提供 (0.0073), 問題 (0.0068), 情報 (0.0060), 師 (0.0055), 知 (0.0050), ナレッジ (0.0050), 解決 (0.0044),
トピック 4 リエンジニアリング	リエンジニアリング (0.012), システム (0.0094), 部門 (0.0083), 経営 (0.0072), 手法 (0.0068), 年 (0.0068), 成功 (0.0063), 事例 (0.0063), 解説 (0.0057), プロセス (0.0055)

データを LDA に適用し，各トピックに出現する単語を出現確率順に表 3.2 に並べた．表 3.2 より，各トピックは，トピック 1 は科学的管理論，トピック 2 は産業革命，トピック 3 はナレッジマネジメント，トピック 4 はリエンジニアリングと解釈した．これらは授業の中で扱われた重要なキーワードでもあり，この授業でのレポートのトピックがこれらによって構成されることには妥当性がある．各レポートは，この 4 つのトピックを組み合わせて書かれており，それぞれのトピックの重みを示すトピック分布がレポートの主題を反映している．したがって，トピック分布が類似した 2 つのレポートは，それぞれの主題も類似していると解釈できた．つまり，式（3.15）を用いて各レポート同士のトピック分布の距離を算出することにより，レポートの主題を同定することができる．

3.5 レポート推薦システム

本節では，LDA を用いて学習者に過去のレポートを推薦するシステムを提案する．教育分野における従来の推薦システムは，学習者データに表層的に類似したコンテンツ，人，メッセージを推薦していた．しかし，レポートの推薦においては，文字列などが似たレポートを推薦するよりも，同一主題であるが，できるだけ似ていない内容で推薦する方がレポートの書き方に対して深い学習ができると考えられる．主題が同じである 2 つのレポートの単語分布が異なるほど，それらの「構成」,「表現」,「オリジナリティ」が異なる確率が高まる．そのために，トピック分布が類似で異なる単語分布のレポートを推薦することで，多様な「構成」,「表現」,「オリジナリティ」を持つレポートを推薦できると考えられる．したがって，本論文では，LDA を用いてレポートの主題を同定し，学習者のレポートに用いられる文字列とはなるべく異なるレポートを推薦するシステムを提案する．

3.5.1 推薦メカニズム

本論文で提案するレポート推薦システムの概要は以下のとおりである．

> 学習者が入力したレポートとトピック分布の距離（式（3.15））が小さい順に N 個の他者レポートをデータベースから抽出する．
> その中から，学習者が入力したレポートと使用単語の距離（式（3.16））が大きい順に M 個の他者レポートを抽出する．
> その中から，ピアレビューによるレポートの得点が高い順に S 個の他者レポートを学習者に提示する．

提案手法のアルゴリズムの擬似コードを Algorithm1 に示す．Algorithm1 の 2 行目は，あらかじめ他者のレポート集合 $Y = \{d_1, d_2, \ldots, d_i, \ldots, d_n\}$ を LDA により分析し，それぞれ他者レポートのトピック分布 θ_{d_i}， ϕ を推定す

Algorithm 1 提案手法の擬似コード

1: input: d_x : 学習者のレポート

$Y = \{d_1, d_2, \ldots, d_i, \ldots, d_n\}$: 他者のレポート集合

output : Z :推薦レポート集合

2: あらかじめ他者のレポート集合 Y を LDA により分析し，

それぞれ他者レポートのトピック分布 θ_{d_i}, ϕ を推定する．

3: 推定した ϕ を用いて，d_i のトピック分布 θ_{d_x} を計算する．

4: 式（3.15）を用い d_x と Y 中のそれぞれの他者レポートのトピック分布 θ_{d_i} 間の距離 $T_{\mathrm{JSD}}(d_x, d_i)$ が小さな N 個を集合 Z とする．

5: 式（3.16）を用い, 学習者のレポート d_x と Z 中のそれぞれ他者レポート d_i の単語分布の距離 $W_{\mathrm{JSD}}(d_x, d_i)$ が大きな M 個を集合 Z とする．

6: Z 中の他者のレポートからレポート評価が高い S 個を集合 Z とする．

7: **return** Z

る．3 行目は，θ_{d_i} を用いて，学習者が作成したレポート d_x のトピック分布 θ_{d_x} を推定する．4 行目は，式（3.15）を用い d_x と Y 中のそれぞれの他者レポートのトピック分布 θ_{d_i} 間の距離 $T_{\mathrm{JSD}}(d_x, d_i)$ が小さな順に N 個のレポートを集合 Z として抽出する．5 行目は，式（3.16）を用い, 学習者のレポート d_x と Z 中のそれぞれ他者レポート d_i の単語分布の距離 $W_{\mathrm{JSD}}(d_x, d_i)$ が大きな M 個を集合 Z とする．6 行目は，Z 中の他者のレポートからレポート評価が高い S 個を集合 Z とする．また，本論文では $N = 15$，$M = 10$，$S = 4$ とした．$N, M, S(N > M > S)$ の各値は，各主題のレポートを推薦するため，各主題のレポート数の平均値である 22.5 を上限値とした．ただし，トピック分布 θ_{d_i} の推定には，十分な単語数・語彙数が必要であることが知られているが，レポートなどの比較的長文の解析では問題がない [27].

図 3.5.1 レポート推薦画面

3.5.2 本推薦システムの推薦画面

ここでは，実装した本推薦システムのインタフェースについて述べる．まず，学習者が作成したレポートをシステムに入力する．次にシステムは図 3.5.1 のようにレポートを推薦する．上部は，推薦レポートが並び，見たいレポートを選択して閲覧できる．詳細情報には，レポートの講義名，書き出し，推薦レポートのトピック分布のグラフが表示される．また，システムは入力したレ

ポートの統計情報も表示する．ここでは，トピック分布，出現単語のランキング，トピックに属する単語を確率順に並べて表示する．推薦レポートと入力したレポートのトピック分布を比較することで，どのトピックが類似し，そのトピックに含まれる単語群がわかる．

3.6 評価

3.6.1 実験

本推薦システムを用いたレポート作成支援の被験者実験について述べる．本論文では，同一主題でなるべく内容（用いられる単語分布）が似ていないレポートを推薦することが学習に効果的であると仮定している．しかし，その対立仮説として，同一主題でなるべく内容が似ているレポートの推薦，異なる主題でなるべく内容が似ているレポートの推薦，異なる主題でなるべく内容が似ていない推薦が考えられる．トピック分布，単語分布ともに類似のレポートを推薦する場合には，類似の論文ばかりが推薦され，多様性も少ないと予想できる．トピック分布の類似度が低い場合は，単語分布の類似度に関わらず，学習者が深く考えている主題と無関係になってしまい，学習効率が悪くなると予想できる．本実験では，本論文の仮説に対応する実験群を B，対立仮説に対応する統制群として A，C，D を導入した．トピック分布の類似度，単語分布の類似度による効果を確認するために，ランダムに推薦する実験群 R を設け，以下のように実験を設定した．

A 群： トピックの類似度が高いレポートの中から，使用している単語の類似度が高いものを推薦する．

B 群（提案手法）： トピックの類似度が高いレポートの中から，使用している単語の類似度が低いものを推薦する．

C 群： トピックの類似度が低いレポートの中から，使用している単語の類似度が高いものを推薦する．

D 群： トピックの類似度が低いレポートの中から，使用している単語の類似度が低いものを推薦する．

E 群： TFIDF の類似度が高いものを推薦する．

R 群： ランダムに推薦する．

また，従来からよく用いられてきた文章推薦手法の一つとして TFIDF 値のコサイン類似度の高いものを推薦する手法が知られている（例えば [16, 25]）．この推薦手法をベースラインの統制群として E 群を導入した．各統制群に被験者をランダムに割り振った．各群の被験者数は，A 群は 7 人，B 群は 5 人，C 群は 7 人，D 群は 6 人，E 群は 6 人，R 群は 5 人とした．

まず，被験者全員に同じ課題（「企業における従来の知識創産手法とその問題点について述べよ」）のレポートを作成させた．参考資料として，講義で使用された資料を提供し，文献やインターネットの利用に制限は設けなかった．以下，このときに実験被験者が作成したレポートを事前レポートと呼ぶこととする．次に，事前レポートを本システムに入力後，システムが推薦するレポートを読み，事前レポートを修正してもらい，提出してもらった．修正後のレポートを事後レポートと呼ぶことにする．実験に使用した過去のレポートは，第 3.4.2 節のものと同一である．また，事前レポート，事後レポートについてそれぞれ 2 人の専門家により，以下に示す 5 段階尺度を用いて合議による評価をするように依頼した． 5 段階尺度は，1. 全く思わない，2. やや思わない，3. どちらともいえない，4. やや思う，5. 強く思う，とした．評価項目を表 3.3 に示す．

表 3.3 レポートの評価項目

	評価内容
評価項目 1	レポートの構成は良かったですか.
評価項目 2	レポートの表現方法はよかったですか.
評価項目 3	レポートのオリジナリティは良かったですか.

ただし，評価項目 1 では，レポートの主張点の論理構成や文章の流れの良さを評価するように専門家に依頼した．

3.6.2 実験結果

本節では，被験者実験の結果を考察する．事前レポートと事後レポートの専門家による評価結果を，表 3.4, 表 3.5 に示す．各表の数値は評価値の平均と分散（カッコ内）を示す．事前レポートと事後レポートの評価結果に対して各評価項目に分散分析を行い，各推薦手法に割り当てられた学習者のレポート作成技術の差について分析する．分散分析には，Kruskal-Wallis 法および一元配置分散分析法を用いた．分散分析の帰無仮説は，「各推薦手法間で差がない」，対立仮説は，「各推薦手法間で差がある」とした．Kruskal-Wallis 法は，正規分布を仮定していない 3 群以上の各群の平均の差の検定を行う手法である．事前レポートと事後レポートに対して，各評価項目にそれぞれ分散分析を行った結果を表 3.4, 表 3.5 に示す．各表の数値は評価値の平均と分散（カッコ内）を示す．分散分析における自由度は 4 であった．各表の p 値は，各評価項目における分散分析の p 値を示す．

表 3.4 事前レポートの評価結果：平均と分散（カッコ内），分散分析結果

推薦手法（被験者数）	構成	表現	オリジナリティ
A(7)	2.29(0.49)	2.43(0.245)	2.0(0.0)
B(5)	2.6(0.24)	2.8(0.16)	2.2(0.16)
C(7)	2.86(0.69)	2.86(0.41)	2.29(0.2)
D(6)	2.5(0.25)	2.67(0.222)	2.17(0.14)
E(6)	2.667(0.222)	2.5(0.25)	2.0(0.0)
R(5)	2.2(0.16)	2.2(0.16)	2.2(0.16)
P 値（Kruskal-Wallis）	0.61	0.36	0.61
P 値 (一元配置分散分析)	0.63	0.59	0.48
分散比 F 値	0.649	0.706	0.90

表 3.4 より，事前レポートは，すべての評価項目において，帰無仮説が棄

表 3.5　事後レポートの評価結果：平均と分散（カッコ内），分散分析結果

推薦手法（被験者数）	構成	表現	オリジナリティ
A(7)	2.43(0.531)	2.71(0.49)	2.0(0.0)
B(5)	4.2(0.16)	4.0(0.0)	3.2(0.16)
C(7)	2.14(0.98)	2.86(0.408)	1.86(0.408)
D(6)	2.5(0.25)	2.67(0.222)	2.17(0.139)
E(6)	2.67(0.222)	3.0(0.0)	2.17(0.139)
R(5)	2.2(0.16)	2.2(0.16)	2.2(0.16)
P 値（Kruskal-Wallis）	0.0094	0.0037	0.0045
P 値 (一元配置分散分析)	0.00092	0.0025	0.00039
分散比 F 値	6.49	5.44	7.44

表 3.6　事前，事後レポートと推薦レポートの単語数の平均値と分散（カッコ内）

	単語数		
推薦手法（被験者数）	事前レポート	事後レポート	推薦レポート
A(7)	418.57(14903.39)	433.14(14924.41)	299.24(10822.44)
B(5)	446.2(17114.96)	512.4(25703.04)	467.75(9424.62)
C(7)	392.86(11871.84)	414.14(14978.12)	230.78(7510.34)
D(6)	459.67(8138.89)	496.83(5531.14)	387.42(9800.81)
E(6)	414.3(7777.2)	442.7(11285.2)	463.18(42164.59)
R(5)	289 (6762.4)	293.4(7800.24)	273.2 (17911.3)

却されず，有意差は認められなかった．表 3.5 より，事後レポートは，すべての評価項目において，帰無仮説が棄却され，各推薦手法間で差があることを確認したため，各評価項目に Steel-Dwass 法を用いて多重比較を行い，以下のことがわかった．

「構成」では，多重比較の結果，提案手法 B が，手法 A に対して有意水準 5%（p 値 0.037），手法 C に対して有意水準 10%（p 値 0.08），手法 D に対して有意水準 5%（p 値 0.048），手法 E に対して有意水準 5%（p 値 0.046），手法

表 3.7　事前，事後レポートと推薦レポートの語彙数の平均値と分散（カッコ内）

	語彙数		
推薦手法（被験者数）	事前レポート	事後レポート	推薦レポート
A(7)	240.57(3143.39)	245.71(3208.49)	190.73(3585.27)
B(5)	231.2(2346.96)	252.6(1605.84)	289.65(4202.98)
C(7)	214.14(3073.55)	223.57(4178.95)	147.65(7510.34)
D(6)	242.33(1399.89)	260.83(1211.47)	230.19(3515.84)
E(6)	230.7(2187.6)	241.0(2748.0)	269.06(8333.25)
R(5)	177 (2149.2)	182.2 (2233.4)	172.9 (4573.2)

表 3.8　修正文章数

推薦手法（被験者数）	修正文章数
A(7)	21
B(5)	74
C(7)	19
D(6)	17
E(6)	28
R(5)	5

R に対して有意水準 10%（p 値 0.059）で有意差があった．他手法間での有意差は認められなかった．提案手法 B は，トピックの類似度が高く，出現単語の類似度が低いレポートを推薦する手法である．同一主題で，なるべく内容（単語分布）が互いに異なるレポートを推薦するため，同一主題で自身のレポートと異なる多様な他者のレポート構成から学習できたと考える．

「表現」では，多重比較の結果，提案手法 B が，手法 A に対して有意水準 10%（p 値 0.087），C に対して有意水準 10%（p 値 0.083），D に対して有意水準 5%（p 値 0.037），E に対して有意水準 5%（p 値 0.019），R に対して有意水準 5%（p 値 0.045）で手法で有意差があった．

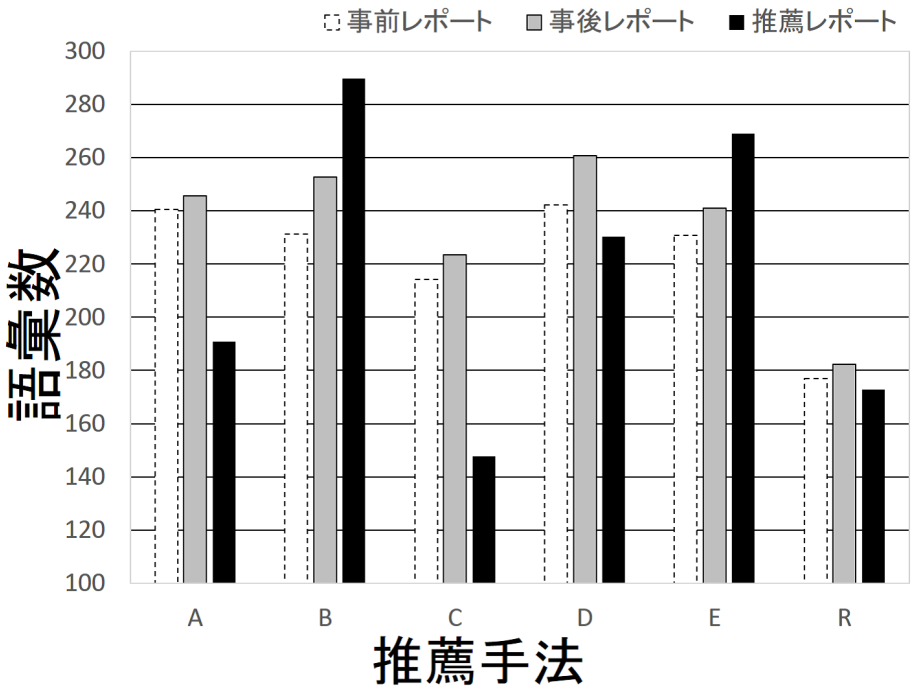

図 3.6.1　レポートの単語数

他手法間での有意差は認められなかった．提案手法 B により推薦されるレポートは，類似した主題で，内容が互いに異なるため，多様な他者の表現方法を学習できたと考える．

「オリジナリティ」では，多重比較の結果，提案手法 B が，A に対して有意水準 5%（p 値 0.015），C に対して有意水準 10%（p 値 0.082），D に対して有意水準 10%（p 値 0.09），E に対して有意水準 10%（p 値 0.091）R に対して有意水準 5%（p 値 0.045）で手法で有意差があった．

他手法間での有意性は認められなかった．提案手法 B では，なるべく内容が互いに異なるレポートが推薦され，個々のオリジナリティを学ぶことができたと考える．

事前レポート，事後レポート，推薦レポートについて，単語数，語彙数のグループ内平均と分散を表 3.6，表 3.7 に示し，単語数，語彙数の平均値を図

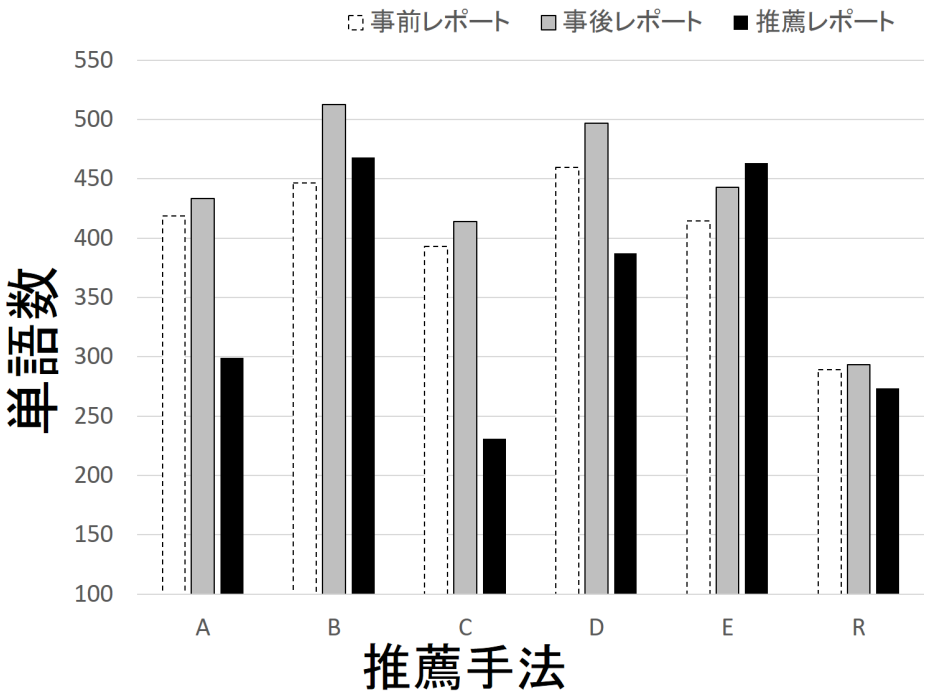

図 3.6.2　レポートの語彙数

3.6.1，図 3.6.2 に示した．また，各推薦手法の事前レポートと事後レポートを比較し，学習者が修正したと著者らが判断した箇所（修正文章）の総数を数え上げ，表 3.8 に追加した．この結果から，各推薦手法で以下のことが分かった．

A：　手法 A は，トピックの類似度が高く（主題が似ている），出現単語の類似度が高いレポートを推薦する手法である．手法 A によって推薦されるレポートは，学習者自身が書いたレポートと内容が類似するため，自身のレポートとの差異が少なすぎて，修正箇所が少ないと考えられる．

B：　手法 B は，トピックの類似度が高く，出現単語の類似度が低いレポートを推薦する手法である．推薦レポートの単語数や語彙数が最も多かった．事前レポートと事後レポートの単語数や語彙数の変化量，修正文章数が最も多かった．トピック分布が類似しているのに，学習者のレポートとの単語分布の差を最大にしようとするために，類似した主題である

が，内容が互いに異なるため，自身のレポートとは異なる多様な他者のアプローチや例から学習できたと考えられる．

C： 手法 C は，トピックの類似度が低く（主題が異なり），出現単語の類似度が高いレポートを推薦する手法である．手法 C は単語数，語彙数の変化量と修正が少なかった．また，推薦レポートの単語数，語彙数が最も少ないこともわかる．異なる主題でのレポート推薦では，単語が似ていても多様な他者からの学習が促進されないことがわかる．

D： 手法 D は，トピックの類似度が低く，出現単語の類似度が低いレポートを推薦する手法である．手法 D は修正箇所が少なかった．異なる主題であり，表現や例が異なるレポートは，自身のレポートとの内容が大きく異なり，差異が大きすぎて，修正が少なくなったと考える．

E： 手法 E は，TFIDF 値が高いものを推薦する手法である．手法 E は修正箇所が少なかった．手法 E によって推薦されるレポートも，学習者のレポートとの差異が少なすぎて，修正箇所が少ないと考えられる．

R： 手法 R は，ランダムに推薦する手法である．手法 R では修正箇所が少なかった．手法 R によって推薦されるレポートは，学習者のレポートとのトピック分布および単語分布を考慮していない．そのため，推薦レポートを読むことにより，他者からの学びにならず，修正箇所が少ないと考えられる．

次に，設定どおりに推薦が行えているかを確認するため，各推薦手法に，事前レポートと推薦されたレポートとのトピック分布の非類似度を式（3.15）を用いて算出し，平均値を図 3.6.3 に示す．単語分布の非類似度を式（3.16）を用いて算出し，平均値を図 3.6.4 に示した．図 3.6.3 より，手法 A，B はトピック分布の非類似度が小さいレポートを推薦する手法であり，手法 C，D はトピック分布の非類似度が大きくなるレポートを推薦する手法であることが確認できる．図 3.6.4 より，手法 A，C は単語分布の非類似度が小さいレポートを推薦する手法であり，手法 B，D は単語分布の非類似度が大きくなるレポート

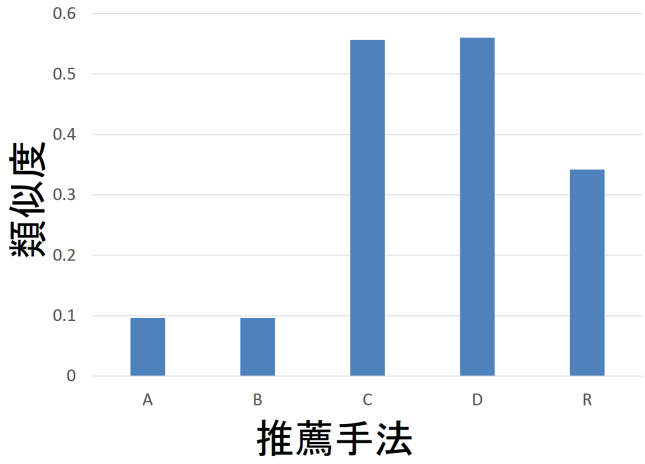

図 3.6.3　事前レポートと推薦されたレポートのトピック分布の非類似度

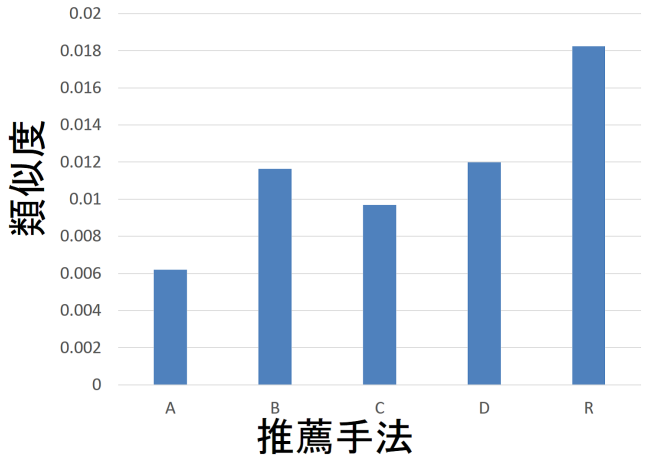

図 3.6.4　事前レポートと推薦されたレポートの単語分布の非類似度

を推薦する手法であることが確認できる．手法 R はランダムに推薦する手法である．

3.6.3 アンケート調査

実験終了後，被験者に表 3.9 に示す質問項目でアンケート調査を行った．アンケートへの回答は，1．全く思わない，2．やや思わない，3．どちらともいえない，4．やや思う，5．強く思う，の 5 段階尺度を用いた．

表 3.9 アンケート調査の質問項目

	質問内容
質問 1	推薦レポートは，主題との関連性があった．
質問 2	推薦レポートを読むことで，通常のレポート作成時に比べ，レポートを見直し改善できた．
質問 3	推薦レポートの全体の構成が参考になった．
質問 4	推薦レポートの文章の表現方法が参考になった．
質問 5	推薦レポートを読むことで，教科書や web だけを勉強していてはわからない知識を得ることができた．

表 3.10 に，アンケート結果を示す．各表の数値は評価値の平均と分散（カッコ内）を示す．得られたアンケート結果に対し，各質問項目ごとに Kruskal-Wallis 法及び一元配置分散分析法を用いて分散分析を行い，表 3.10 に p 値，分散比 F 値を示す．自由度は 4 であった．分散分析から，有意差の認められたアンケート項目について，Steel-Dwass 法を用いて多重比較を行い，以下の結果を得た．

> 質問 1 は，推薦レポートと主題との関連性に関する質問であり，分散分析から有意水準 1% で有意差があり，各推薦手法間に差があることがわかった．多重比較の結果，有意差は示せなかったが，手法 A，B の評価値が高かった．手法 A および提案手法 B は，類似した主題のレポートを推薦する手法である．手法 C，D は，主題が似ていないレポートを推薦する手法であり，手法 R はランダムに推薦する手法である．そのため手法 A，B の評価値が高かったと考えられる．

表 3.10　アンケート結果

推薦手法（被験者数）	質問 1	質問 2	質問 3	質問 4	質問 5
A(7)	4.29 (1.06)	3.86(1.55)	3.86(1.84)	2.86(0.98)	3.71(1.63)
B(5)	4.2 (0.16)	4.4(0.24)	4.4(1.44)	4.6(0.24)	3.8(1.36)
C(7)	2.57 (1.67)	3.0(0.86)	2.57(1.67)	3.29(1.35)	4.0(1.14)
D(6)	2.17 (1.14)	3.33(0.89)	3.83(0.81)	2.33(.89)	4.17(1.14)
E(6)	4.0 (1.0)	3.67(0.89)	4.33(0.22)	4.17(0.47)	3.83(2.81)
R(5)	3.0 (0.5)	3.8(1.2)	3.0(1.0)	2(0.5)	3.4(3.3)
p 値 (Kruskal-Wallis)	0.0081	0.199	0.048	0.0023	0.96
p 値 (一元配置分散分析)	0.0017	0.32	0.066	0.004	0.979
分散比 F 値	4.40	1.21	2.34	6.07	0.18

質問 2 は，推敲に関する質問である．分散分析から有意差は示せなかったが，提案手法 B の評価値が最も高かった．第 3.6.2 節の実験で，手法 B が最も多くのレポート修正を誘発しており，被験者自身も自覚していることがわかる．

質問 3 は，推薦レポートの構成に関する質問である．分散分析から有意差は示せなかったが，提案手法 B の評価値が最も高かった．　第 3.6.2 節の実験でも手法 B がレポートの構成の改善に効果があることがわっており，被験者自身も認識できたと考える．

質問 4 は，推薦レポートの表現に関する質問である．分散分析から有意水準 1% で有意差があった．多重比較の結果，提案手法 B が，手法 A に対して有意水準 10%（0.09），手法 D に対して有意水準 10%（p 値 0.09），手法 R に有意水準 10% で有意差があった．手法 A は，類似した主題で，表層的な単語が類似したレポートを推薦する手法である．自身のレポートとの差異が少なすぎて，表現が参考にならなかったと考えられる．手法 D は，異なる主題で，なるべく内容が互いに異なるレポートを推薦する手法である．自身のレポートと全く関係のないレポートのため，表現が参考にならなかったと考えられる．第 3.6.2 節の解釈を裏

付ける結果となった．

質問 5 は，分散分析から有意差は示せなかったが，手法 C，D の評価値が高かった．手法 C，D は，異なる主題のレポートを推薦する手法である．推薦レポートから，異なる主題の知識を学習したため，手法 C，D の評価値が高くなったと考えられる．

3.7 むすび

レポートライティングにおける他者からの学びを支援するために，過去の学ぶべきレポートを学習者に推薦するシステムを提案した．その特徴は，(1) LDA により，学習者のレポートの潜在的なトピックを推定し，他者レポートとのトピック分布の距離を計算して，同一の主題を扱う他者レポートを検索する手法を提案した．さらに，(2) 学習者のレポートと他者レポートとの単語分布の距離を計算し，同一の主題を扱うが，内容（用いられる単語分布）の異なる評価の高い他者のレポートを多様に推薦する手法を提案したことである．

本システムの有効性を示すため，実際の理工系大学生を対象に評価実験を行った．その結果，提案手法を用いると，簡単には習得できないスキル，レポートの構成，表現，オリジナリティの改善が見られた．また，アンケートにより，提案手法の有効性を示した．

本論文では，他者からの学びによるレポートライティングの学習支援を提案したが，学習者の学び方の能力を向上させたかは示せていないため，今後の課題とする．

第 4 章

LDA におけるトピック数の推定

4.1　はじめに

LDA を用いる場合，トピック数 K を事前に決定しておく必要がある [27]．第 3 章ではトピック数を人手の分類を基に決定している．しかし，データが大量になった場合や新たにデータを追加する際に人手による分類を作成しなおす必要があり，システムを利用する上で現実的ではない．また，人手による分類に即したトピック数が，モデルの学習・推定精度を高くする保証はない．

これまでにトピック数する手法として次のようなものが挙げられる．周辺尤度最大化による決定方法 [36, 41]，アルゴリズムによる決定方法 [42]，第 3 章で用いたような人手による分類結果を用いる方法である．周辺尤度最大化は，トピック数を変え周辺尤度を計算し，周辺尤度の値が最も高くなるときのトピック数をモデルの真のトピック数として採用する方法である．アルゴリズムによる決定方法は，LDA モデルを拡張することで実現している．

本章では，LDA のトピック数推定について議論するため，周辺尤度最大化を用いる．しかし，トピック数を推定する際に用いる周辺尤度を解析的に直接計算できない [41, 43, 44]．そのため，これまでに周辺尤度の近似手法が提案されてきた [36, 41, 43–45]．周辺尤度最大化によりトピック数を推定するとき，トピック数の推定値が LDA のハイパーパラメータの値に影響されることがモ

デルから考えられるが，これまでハイパーパラメータとトピック数の推定について十分に議論されていない．

Griffiths と Steyvers(2004) [36] らは，周辺尤度を調和平均に近似している．また，Griffiths と Steyvers(2004) [36] らは，トピックと語彙に影響するハイパーパラメータ β が大きくなればトピック数は過小評価され，小さくなればトピック数は過大評価されると言及しているが，十分な議論はされていない．本章で，Griffiths と Steyvers(2004) [36] らが提案した周辺尤度を用いてトピック数推定のシミュレーションを行ったところ，Griffiths と Steyvers(2004) [36] らの言及とシミュレーション結果が一致した．

Taddy [41] は，ラプラス近似を用いて以下の周辺尤度を近似している．ハイパーパラメータ 1.0 として，シミュレーションによりトピック数を推定している．しかし，ハイパーパラメータ 1.0 とする根拠は述べられていない．また，ハイパーパラメータの値を変化させたときのトピック数の推定値については言及されていない．本章で Taddy [41] が提案した周辺尤度を用いてトピック数推定のシミュレーションを行ったところ Griffiths と Steyvers(2004) [36] らの言及とは逆の現象を確認した．

以上より，これまで周辺尤度の近似手法が提案されてきたが，ハイパーパラメータの値とトピック数の推定値の関係性について十分な議論がされていない．

本章では，シミュレーションにより，LDA のハイパーパラメータがトピック数の推定値に大きく影響することを示す．次に，漸近解析によりハイパーパラメータが 1 としたときの周辺尤度を最大化することにより，LDA のトピック数を最も正確に推定できることを提案する．この結果を本推薦システムに組み込むことで，その有効性を示す．

4.2 トピック数推定における関連研究

これまでに LDA のトピック数の推定は，perplexity（単語の平均分岐数）最小化 [27]，周辺尤度を最大化する手法 [36, 41, 43–45] が用いられてきた．本節では，各手法を紹介する．

4.2.1 perplexity 最小化によるトピック数の推定

perplexity は，言語モデルの学習精度を測る指標として用いられ，単語の平均分岐数（ある単語の次に続く確率が高い単語の数）を示す [27]．値が小さいほどモデルの分類性能が高いことを示し，以下のように表せる [27]．

$$\begin{aligned} perplexity(w) &= exp\left(\frac{\sum_{d=1}^{D}\log p(w_d)}{\sum_{d=1}^{D}N_d}\right) \\ &= exp\left(\frac{\sum_{d=1}^{D}\sum_{v=1}^{V}\log\sum_{k}^{K}\theta_{dk}\phi_{kv}}{\sum_{d=1}^{D}N_d}\right) \end{aligned} \tag{4.1}$$

トピック数を変え，最小値をとるときのトピック数をモデルのトピック数とする方法が用いられる [27]．

4.2.2 周辺尤度最大化によるトピック数の推定

LDA のトピック数を推定することは，モデル選択と同義である．周辺尤度をトピック数 K を変え計算し，最大値をとったときのトピック数をモデルのトピック数とする手法である [36, 41, 44]．データ W，トピック Z，トピック数 K としたとき，LDA の周辺尤度 $p(W \mid K, \alpha, \beta)$ は以下のように表せる．

$$p(W \mid K, \alpha, \beta) = \sum_{Z} p(W, Z \mid K, \alpha, \beta) = \sum_{Z} p(W \mid Z, K, \beta)p(Z \mid K, \alpha) \tag{4.2}$$

しかし，全ての文書中の単語に対してすべてのトピックのパターンを考慮する必要があり，解析的に計算出来ない．そこで，z^s を $p(Z \mid K, \alpha)$ からの s

番目のサンプル，S をサンプリング数として，以下のように近似できる．

$$p(W \mid K, \alpha, \beta) = \frac{1}{S} \sum_{s} p(W \mid z^s, K, \beta) \tag{4.3}$$

しかし，この手法では，計算効率が悪いことが指摘されている [44].

調和平均による周辺尤度

Griffiths と Steyvers(2004) [36] は，LDA の学習アルゴリズムに崩壊型ギブスサンプリングを提案しており，この学習アルゴリズムを用いることで，周辺尤度を $p(W \mid Z, \beta, K)$ の調和平均 (Harmonic Mean,HM) に近似している．そのため LDA の学習アルゴリズムに崩壊型ギブスサンプリングを用いる場合，この調和平均トピック数推定が用いられてきた．

崩壊型ギブスサンプリングのサンプリング数を S とするとき，Griffiths と Steyvers [36] の周辺尤度は，以下のように導ける [36, 46].

$$p(w \mid K) = \sum_{s=1}^{S} p(w \mid z^s, \beta, K) p(z^s \mid \alpha, K) \tag{4.4}$$

$$= \frac{\sum_{s=1}^{S} p(w \mid z^s, \beta, K) \frac{p(z^s \mid \alpha, K)}{p(z^s \mid w, \alpha, K)}}{\sum_{s=1}^{S} \frac{p(z^s \mid \alpha, K)}{p(z^s \mid w, \alpha, K)}}$$

ここで,

$$p(z^s \mid w, \alpha, K) = \frac{p(w \mid z^s, \beta, K) p(z^s \mid \alpha, K)}{p(w)}$$

と表せるので,

$$\frac{p(z^s \mid \alpha, K)}{p(z^s \mid w, \alpha, K)} = \frac{p(z^s \mid \alpha, K)}{\frac{p(w \mid z^s, \beta, K) p(z^s \mid w, \alpha, K)}{p(w)}} = \frac{p(w)}{p(w \mid z^s, \beta, K)}$$

以上より，$p(w \mid K)$ は，以下のように表せる．

$$p(w \mid K) = \frac{\sum_{s=1}^{S} p(w \mid z^s, \beta, K) \frac{p(z^s \mid \alpha, K)}{p(z^s \mid w, \alpha, K)}}{\sum_{s=1}^{S} \frac{p(z^s \mid \alpha, K)}{p(z^s \mid w, \alpha, K)}}$$

$$= \frac{\sum_{s=1}^{S} p(w \mid z^s, \beta, K) \frac{p(w)}{p(w \mid z^s, \beta, K)}}{\sum_{s=1}^{S} \frac{p(w)}{p(w \mid z^s, \beta, K)}}$$

$$= \frac{\sum_{s=1}^{S} p(w)}{p(w) \sum_{s=1}^{S} \frac{1}{p(w \mid z^s, \beta, K)}} = \frac{S}{\sum_{s}^{S} 1/p(w \mid z^s, K)} \tag{4.5}$$

この調和平均による周辺尤度は，直接計算すると値が桁落ちしてしまうため，トピック数の推定値を $\hat{K}$ として，以下のように計算することで桁落ちを回避できる．

$$\hat{K} = \underset{K}{\operatorname{argmin}} \sum_{s}^{S} \frac{1}{p(w \mid z^s, K)} = \underset{K}{\operatorname{argmin}} \log \sum_{s}^{S} \frac{1}{p(w \mid z^s, K)}$$

最小値 $p(w \mid z^i, K) = \min_i p(w \mid z^i, K)$ として，

$$\log \sum_{s}^{S} \frac{1}{p(w \mid z^s, K)} = \log \left(\frac{1}{p(w \mid z^i, K)} \left(1 + \sum_{s}^{S} \frac{p(w \mid z^i, K)}{p(w \mid z^s, K)} \right) \right)$$

$$= -\log p(w \mid z^i, K) + \log \left(1 + \sum_{s}^{S} \frac{p(w \mid z^i, K)}{p(w \mid z^s, K)} \right)$$

$$= -\log p(w \mid z^i, K)$$

$$+ \log \left(1 + \sum_{s}^{S} exp\left(\log p(w \mid z^i, K) - \log p(w \mid z^s, K) \right) \right) \tag{4.6}$$

Griffiths と Steyvers(2004) [36] は，実データ（論文誌 PNAS(Proceedings of the National Academy of Sciences) の 11 年分の論文の概要）を用いて，周辺尤度（式（4.5））を最大化することによりトピック数の推定を行っている．このとき，ハイパーパラメータを $\alpha = 50/K, \beta = 0.1$ としているが，その根拠については明記されていない．また，ハイパーパラメータ β がトピック数に影響を与えることを以下のように述べている [36].

- β を大きな値にするとトピック数は少なくなる．
- β を小さな値にするとトピック数は大きくなる．

しかし，α については言及されておらず論文中で十分な議論がされていない．また，推定精度が低いことが指摘されている [44]

Newman らの周辺尤度

Newman ら [47] は，トピック数の決定に関して十分な議論はされていないが，LDA モデルの評価尺度である *perplexity* から周辺尤度を導出している．

周辺尤度は以下の式で表される．

$$p(w \mid K) = \sum_{d=1}^{D}\sum_{v=1}^{V} \log(\frac{1}{S}\sum_{s=1}^{S}\sum_{k=1}^{K} \theta_{dk}^{s}\phi_{kv}^{s}), \tag{4.7}$$

$$\theta_{d,k} = \frac{N_{kd} + \alpha_k}{N_d + \sum_{k=1}^{K}\alpha_k}, \tag{4.8}$$

$$\hat{\phi_{k,v}} = \frac{N_{kv} + \beta_v}{N_k + \sum_{v=1}^{V}\beta_v}. \tag{4.9}$$

ここで S は崩壊型ギブスサンプリングのサンプリング回数である．

ラプラス近似による周辺尤度

Taddy [41] は，ラプラス近似を用いて以下の周辺尤度を導出し，シミュレーションデータを用いてトピック数の推定実験を行っており，高精度にトピック数を推定している．

$$p(W \mid K) \approx p(W, \hat{\Theta}, \hat{\Omega}) \mid -H \mid^{-\frac{1}{2}} (2\pi)^{\frac{d}{2}} K! \tag{4.10}$$

ここで，

$$p(W, \hat{\Theta}, \hat{\Phi}) = \prod_{d=1}^{D} MN(w_d; \hat{\Theta}\hat{\Phi}) p(Z \mid \alpha) \prod_{k=1}^{K} p(W \mid Z, \beta)$$

ここで $d = KV + (K-1)D$，$MN(\cdot)$ は多項分布，H はヘッセ行列を示す．

以上より，これまでトピック数の推定における周辺尤度が提案されてきた．しかし，トピック数の推定において，ハイパーパラメータが大きく影響すると考えられるが，十分な議論がされていない．

本章では，シミュレーションデータを用いて，周辺尤度（式（4.7），式（4.10））最大化により推定する．その際，ハイパーパラメータ α，β の値がトピック数の推定値に及ぼす影響について，シミュレーション及び漸近解析による分析を行う．

4.3 シミュレーションデータのトピック数の推定

4.3.1 シミュレーションデータ

シミュレーションデータの生成について述べる．プログラムは公開されているパッケージを用いた [41]．シミュレーションデータ生成のアルゴリズムの擬似コードを Algorithm2 に示す．アルゴリズム 2 の 3 行目は，$\alpha_k = 1/K$ と

Algorithm 2 シミュレーションデータ生成アルゴリズム

1: input:K：トピック数
   D：文書数
   V：語彙数
   N：文書内の単語数の平均値
   N_d：文書 d の単語数
   output：$W = \{w_d\}_{d=1}^{D}$：文書集合
2: **for** $d = 1, \cdots, D$ **do**
3: $\quad \theta_d \sim Dir(\alpha = 1/K)$
4: $\quad N_d \sim Poisson(d, N)$
5: **end for**
6: **for** $k = 1, \cdots, K$ **do**
7: $\quad \phi_k \sim Dir(\beta = 1/V)$
8: **end for**
9: **for** $d = 1, \cdots, D$ **do**
10: $\quad w_d \sim multinomial(\theta_{d1}\phi_1 + \theta_{d1}\phi_1 + \cdots + \theta_{dK}\phi_K, N_d)$
11: **end for**
12: **return** $W = \{w_d\}_{d=1}^{D}$

して，文書 d のトピック分布 θ を生成する [41]．このとき，各文書のトピック分布が同一にならない様にしている．つまり，各文書で出現しやすいトピック

が異なる．4 行目は，文書 d の単語数 N_d をポアソン分布から生成する．7 行目は，$\beta_v = 1/V$ として，トピック k の単語分布 ϕ を生成する．このとき，各トピックの語彙の分布が異なるようにしている．つまり，各トピックで出現しやすい語彙が異なる．10 行目は，文書 d の単語集合 w_d を $\theta\phi$ の多項分布から，各文書における語彙の出現頻度を算出する．

シミュレーションデータ生成に用いた各パラメータ K^{true}，D，V，N のそれぞれの値は，真のトピック数を $K^{true} = 5, 10, 20, 30$，文書数は，$D = 100, 1000$，語彙数は，$V = 100, 1000, 5000$，文書 d に含まれる単語数は，$N_d = 100, 300, 1000, 10000$ とした．ハイパーパラメータ α，β は，$\alpha = 10^{-4}, 1, 10^4$，$\beta = 10^{-4}, 1, 10^4$ とし，$\alpha_k = \alpha/K$，$\beta_v = \beta/V$ とした．シミュレーションデータ生成に用いる各パラメータで独立に 5 つのシミュレーションデータを生成し，各データに対して α，β を変え，周辺尤度（式（4.5），式（4.10））を最大化することによりトピック数を推定した．推定時のハイパーパラメーは $\alpha_k = \alpha/K$，$\beta_v = \beta/V$ とした．ただし，変化させるトピック数は，シミュレーションデータの真のトピック数 K^{true} のとき 1 から $K^{true} + 20$ まで変化させた．推定に用いた周辺尤度は，ラプラス近似による周辺尤度とした．

4.3.2 シミュレーション結果

α，β の組み合せ毎にトピック数 K を変え，周辺尤度最大化によりトピック数を推定した．シミュレーション結果を表 4.1，表 4.2，表 4.3，表 4.4，表 4.5，表 4.6，表 4.7，表 4.8，表 4.9，表 4.10，表 4.11，表 4.12，表 4.13，表 4.14，表 4.15，表 4.16 に示す．表のタイトルにシミュレーションデータのパラメータ（トピック数 K，文書数 D，語彙数 V，文書 d の単語数 N_d）を示した．例えば，表 4.1 はトピック数 $K^{true} = 10$，文書数 $D = 100$, 語彙数 $V = 100$, 文書内の単語数 $N_d = 100$ のシミュレーションデータに対して，ラプラス近似による周辺尤度を用いた推定結果を示す．表中の数値は，トピック数推定に用いた α，β，そのときの平均二乗誤差（mse），トピック数の推定値の平均値（ave），対数周辺尤度（logML）を示す．mse は値が小さいほど推定精度が高

いことを示す．mse は，シミュレーションデータの真のトピック数を K^{true}，推定値を $\hat{K}$，データ数 n とするとき式（4.11）で表せる．

$$mse = \frac{1}{n}\sum_{i=1}^{n}(K^{true} - \hat{K})^2 \tag{4.11}$$

表 4.1　$K^{true} = 10, D = 100, V = 100, N_d = 100$

α	β	mse	ave	logML
0.0001	0.0001	64	2	-4906.38
0.0001	1	56.7	2.5	-2162.14
0.0001	10000	1.6	10	2745.2
1	0.0001	20.1	5.7	2236.22
1	1	16.2	6.2	2572.89
1	10000	20.8	5.6	2161.72
10000	0.0001	64	2	-8213565
10000	1	64	2	-8211382
10000	10000	227.6	24.9	-3275893

表 4.1，表 4.2，表 4.3，表 4.4 は，シミュレーションデータのパラメータが $K^{true} = 10, D = 100, V = 100, N_d = 100, 300, 1000, 10000$ の時の結果を示す．結果からトピック数の推定値は α，β が小さい時は過小評価され，大きい時は過大評価されている．$\alpha = \beta = 1$ のとき真値に近い値を推定している．N_d が大きくなるに従い α，β の値によらず真値に近づくが，$\alpha = \beta = 1$ としたときが最も精度が高い．$\alpha = \beta = 1$ ではないとき，真値に近い値を推定している場合もあるが，対数周辺尤度の値が $\alpha = \beta = 1$ のときの方が高くモデルとして適当であると考えられる．

表 4.5，表 4.6，表 4.7，表 4.8 は，シミュレーションデータのパラメータが $K^{true} = 10, D = 1000, V = 100, N_d = 100, 300, 1000, 10000$ の時の結果を示す．表 4.1，表 4.2，表 4.3，表 4.4 と比べて，文書数 D が十分に大きい．トピック数の推定値は $\alpha = \beta = 1$ のとき真値に近い値を推定し，対数周辺尤度の

表 4.2 $K^{true} = 10, D = 100, V = 100, N_d = 300$

α	β	mse	ave	logML
0.0001	0.0001	4.4	8.2	8357.92
0.0001	1	1	9.6	17279.34
0.0001	10000	2.2	11.2	20288.3
1	0.0001	0.4	9.8	24032.58
1	1	0.5	10.3	26978.45
1	10000	0.3	10.1	25647.14
10000	0.0001	62.13	2.13	-8213924
10000	1	62.13	2.13	-8211476
10000	10000	390.25	29.75	-3763964

表 4.3 $K^{true} = 10, D = 100, V = 100, N_d = 1000$

α	β	mse	ave	logML
0.0001	0.0001	0.4	10.4	96707.16
0.0001	1	0.8	10.6	102868.8
0.0001	10000	36.1	14.7	95841.09
1	0.0001	0.9	10.3	114804.5
1	1	0.2	10.2	117240.8
1	10000	0.3	10.3	117006.4
10000	0.0001	60.5	2.25	-8213058
10000	1	0.88	9.63	-8206974
10000	10000	357.88	28.88	-6864654

値も大きい．α，β が小さい時は，データが十分に大きいため真値に近い値を示しているが対数周辺尤度の値は $\alpha = \beta = 1$ のときよりも小さい．α，β が大きい時は過大評価されている．N_d が大きくなるに従い真値に近い値を推定している場合もあるが，$\alpha = \beta = 1$ としたときが最も精度が高く，周辺尤度の値

表 4.4 $K^{true} = 10, D = 100, V = 100, N_d = 10000$

α	β	mse	ave	logML
0.0001	0.0001	5.5	12.3	13000000
0.0001	1	10.5	13.1	12900000
0.0001	10000	158.5	20.9	12800000
1	0.0001	6	12.4	12700000
1	1	4.7	11.9	12600000
1	10000	5.2	12.2	12900000
10000	0.0001	171.13	22.63	3243604
10000	1	199.75	23.5	3195602
10000	10000	234.88	24.88	804560.7

表 4.5 $K^{true} = 10, D = 1000, V = 100, N_d = 100$

α	β	mse	ave	logML
0.0001	0.0001	2.3	10.9	105761.8
0.0001	1	1.2	10.8	120535.5
0.0001	10000	12.1	13.3	125085.5
1	0.0001	11.2	13.2	123559.6
1	1	7.9	12.7	117068.4
1	10000	12.2	13.2	122459.8
10000	0.0001	400	30	-82000000
10000	1	400	30	-82000000
10000	10000	325.75	28	-33800000

が最も高い.

表 4.9，表 4.10，表 4.11，表 4.12 は，シミュレーションデータのパラメータが $K^{true} = 10, D = 100, V = 1000, N_d = 100, 300, 1000, 10000$ の時の結果を示す．文書数 D, 文書内単語数 N_d に対して語彙数 V が十分に大きく，デー

表 4.6 $K^{true} = 10, D = 1000, V = 100, N_d = 300$

α	β	mse	ave	logML
0.0001	0.0001	8.1	12.5	340487.9
0.0001	1	19.5	14.1	374003.9
0.0001	10000	39.2	16.2	385736.1
1	0.0001	4.9	11.5	386647
1	1	1.2	11	388818
1	10000	7.3	11.7	374803.1
10000	0.0001	400	30	-82000000
10000	1	400	30	-82000000
10000	10000	400	30	-45200000

表 4.7 $K^{true} = 10, D = 1000, V = 100, N_d = 1000$

α	β	mse	ave	logML
0.0001	0.0001	12.8	13.2	1267506
0.0001	1	53.3	16.7	1256186
0.0001	10000	90.5	19.5	1231374
1	0.0001	1.2	11	1271244
1	1	1.1	10.9	1275347
1	10000	1.6	11	1316698
10000	0.0001	380.5	29.5	-81900000
10000	1	400	30	-81900000
10000	10000	213.25	15.5	-82100000

タがスパースなためトピック数を推定することが困難である．N_d を大きくしたとき，$\alpha = \beta = 1$ のとき真値に近い値を推定することができる．

表 4.13，表 4.14，表 4.15，表 4.16 は，シミュレーションデータのパラメータが $K^{true} = 10, D = 1000, V = 1000, N_d = 100, 300, 1000, 10000$ の時の結

表 4.8 $K^{true} = 10, D = 1000, V = 100, N_d = 10000$

α	β	mse	ave	logML
0.0001	0.0001	292.1	26.9	130000000
0.0001	1	344.1	28.5	131000000
0.0001	10000	376.8	29.4	134000000
1	0.0001	18.2	14.2	132000000
1	1	12.7	13.5	130000000
1	10000	19.1	14.1	133000000
10000	0.0001	221.25	23.75	32300000
10000	1	274	26.25	32900000
10000	10000	215	24.43	15800000

表 4.9 $K^{true} = 10, D = 100, V = 1000, N_d = 100$

α	β	mse	ave	logML
0.0001	0.0001	64	2	-77344.8
0.0001	1	64	2	-50309.3
0.0001	10000	64	2	-23591.6
1	0.0001	64	2	-28591.1
1	1	64	2	-28700.8
1	10000	64	2	-28431.4
10000	0.0001	64	2	-8267848
10000	1	64	2	-8240835
10000	10000	385.63	29.63	-3768460

果を示す．文書内単語数 N_d に対して語彙数 V が十分に大きい．表 4.9，表 4.10，表 4.11，表 4.12 と比べて文書数 D が十分に大きい．そのためデータが十分に大きくなり，トピック数の推定値は $\alpha = \beta = 1$ のとき真値に近い値を推定し，対数周辺尤度の値も大きい．α，β が小さい時は，データが十分に大き

表 4.10 $K^{true} = 10, D = 100, V = 1000, N_d = 300$

α	β	mse	ave	logML
0.0001	0.0001	64	2	-73365
0.0001	1	64	2	-45726.2
0.0001	10000	64	2	-20991.9
1	0.0001	64	2	-25719.8
1	1	64	2	-25379.4
1	10000	64	2	-25476.8
10000	0.0001	64	2	-8268732
10000	1	64	2	-8243229
10000	10000	395.13	29.88	-6703947

表 4.11 $K^{true} = 10, D = 100, V = 1000, N_d = 1000$

α	β	mse	ave	logML
0.0001	0.0001	64	2	-60704.6
0.0001	1	64	2	-32335
0.0001	10000	18.4	5.8	11435.27
1	0.0001	42	3.6	-6417.37
1	1	33.5	4.3	-6310.1
1	10000	47	3.2	-6390.9
10000	0.0001	64	2	-8271181
10000	1	64	2	-8243468
10000	10000	54.63	2.63	-8287553

いため真値に近い値を示しているが対数周辺尤度の値は $\alpha = \beta = 1$ のときよりも小さい．α，β が大きい時は過大評価されている．$\alpha = \beta = 1$ としたときが最も精度が高く，周辺尤度の値が最も高い．

これらの結果から，α，β の値にトピック数の推定値が敏感であることが

表 4.12 $K^{true} = 10, D = 100, V = 1000, N_d = 10000$

α	β	mse	ave	logML
0.0001	0.0001	1.4	11	13800000
0.0001	1	4.5	11.9	13900000
0.0001	10000	74	17.2	13400000
1	0.0001	1	10.8	14000000
1	1	1	10.8	13900000
1	10000	0.4	10.4	13800000
10000	0.0001	2.25	11	3770157
10000	1	9.38	11.88	3844194
10000	10000	242	25.25	-1430797

表 4.13 $K^{true} = 10, D = 1000, V = 1000, N_d = 100$

α	β	mse	ave	logML
0.0001	0.0001	64	2	-75804.76
0.0001	1	64	2	-46945.88
0.0001	10000	5.2	7.8	14102.81
1	0.0001	40.1	3.7	809.52
1	1	34	4.2	-3606.75
1	10000	40.1	3.7	-1311.37
10000	0.0001	64	2	-82200000
10000	1	61.86	2.14	-82100000
10000	10000	383.29	29.57	-48000000

わかる．α，β が小さいときトピック数の推定値は過小評価され，逆に大きいとき過大評価される．$\alpha = \beta = 1$ のとき，真のトピック数に近づくと予測できる．ただし，表 4.9，表 4.10，表 4.11，表 4.13 など語彙 V に対して D や N_d が小さい時，つまりスパースなデータのときトピック数の推定精度が低く

表 4.14 $K^{true} = 10, D = 1000, V = 1000, N_d = 300$

α	β	mse	ave	logML
0.0001	0.0001	9.9	7.1	32969.81
0.0001	1	0.8	9.4	141949.51
0.0001	10000	0.2	10.2	272744.83
1	0.0001	0.2	10	234043.7
1	1	0.4	10	230855.47
1	10000	0.4	10.4	234436.75
10000	0.0001	51.43	2.86	-82200000
10000	1	40.29	3.71	-82200000
10000	10000	247.57	20.43	-81900000

表 4.15 $K^{true} = 10, D = 1000, V = 1000, N_d = 1000$

α	β	mse	ave	logML
0.0001	0.0001	0.7	10.5	1006340
0.0001	1	0.2	10.2	1132611.76
0.0001	10000	0.3	10.3	1283951.12
1	0.0001	0.5	10.5	1230648.13
1	1	0	10	1228410.23
1	10000	0.2	10.2	1241345.34
10000	0.0001	50.71	3	-82200000
10000	1	0	10	-82100000
10000	10000	61.86	2.14	-82200000

なる．N_d，D が大きいとき（つまりデータが十分に大きいとき），$\alpha = \beta = 1$ のとき真のトピック数を推定できると考えられる．また，$\alpha = \beta = 1$ のときと同程度の推定精度となる α と β の組み合わせも確認できるが，周辺尤度の値が $\alpha = \beta = 1$ の方が大きいもしくは同程度であるため，$\alpha = \beta = 1$ とするこ

表 4.16 $K^{true} = 10, D = 1000, V = 1000, N_d = 10000$

α	β	mse	ave	logML
0.0001	0.0001	7.8	12.6	141000000
0.0001	1	14.3	13.7	144000000
0.0001	10000	12.2	13.4	141000000
1	0.0001	5.8	12.2	145000000
1	1	6.8	12.4	145000000
1	10000	5.4	12.2	143000000
10000	0.0001	4.86	11.14	42900000
10000	1	5.71	12	42700000
10000	10000	175	21.57	4237934.27

とでトピック数を高精度に推定できると考えられる.

4.4 LDA の周辺尤度の漸近解析

トピック数を推定するために必要な周辺尤度は，以下ののように表せる．

$$p(W \mid K, \alpha, \beta) = \sum_{Z} p(W \mid Z, K, \beta) p(Z \mid K, \alpha) \tag{4.12}$$

この周辺尤度は解析的に解くことが困難であるため，これまで周辺尤度の近似手法が提案されてきた [36, 41, 47]．シミュレーションの結果からトピック数の推定値はハイパーパラメータの値に非常に敏感になる結果を得た．しかし，これまでハイパーパラメータの値とトピック数の関係について十分な議論はされていない [36, 41]．Griffiths と Steyvers [36] らは，β を大きくするとトピック数は過小評価され，β を小さくするとトピック数は過大評価されると述べている．しかし，Taddy(2012) [41] が提案した周辺尤度を用いると Griffiths と Steyvers(2004) ら言及とは逆の現象を確認した．

これらの現象を引き起こす周辺尤度の関係性を明らかにするため，LDA の対数周辺尤度 $\log p(W \mid K, \alpha, \beta)$ の漸近解析を行う．本章の漸近解析は，ベイジアンネットワークにおける対数周辺尤度の漸近解析 [48, 49] を参考にした．

まず式（4.12）の α の項 $p(Z \mid K, \alpha)$，β の項 $p(Z \mid K, \alpha)$ をそれぞれ漸近解析し，トピック数の推定値への影響を総合的に考える．

$$\begin{aligned} \log p(Z \mid \alpha) = D & \left(\log\Gamma \left(\sum_{k=1}^{K} \alpha_k \right) - \sum_{k=1}^{K} \log\Gamma \left(\alpha_k \right) \right) \\ & + \sum_{d=1}^{D} \left(\sum_{k=1}^{K} \log\Gamma \left(N_{kd} + \alpha_k \right) - \log\Gamma \left(N_d + \sum_{k=1}^{K} \alpha_k \right) \right) \end{aligned} \tag{4.13}$$

$$\begin{aligned} \log p(W \mid Z, \beta) = K & \left(\log\Gamma \left(\sum_{v=1}^{V} \beta_v \right) - \sum_{k=1}^{K} \log\Gamma \left(\beta_v \right) \right) \\ & + \sum_{v=1}^{V} \left(\sum_{v=1}^{V} \log\Gamma \left(N_{kv} + \beta_v \right) - \log\Gamma \left(N_k + \sum_{v=1}^{V} \beta_v \right) \right) \end{aligned} \tag{4.14}$$

それぞれの式において事前分布の項（最初の項）と尤度の項（二つ目の項）の

二つ項に分けて分析できる．事前分布の項はハイパーパラメータのみを主な変数として記述されているためデータに依存せず，尤度の項はハイパーパラメータとデータを反映している．

4.4.1 事前分布項の分析

$\log p(Z \mid \alpha)$ の事前分布の項を，α の値により場合分けをして漸近解析する．

定理 1. $\alpha \leq 1.0$ のとき，$p(Z \mid \alpha)$ の事前分布の項は，以下のように漸近展開できる．

$$D\left(\log\Gamma\left(\sum_{k=1}^{K}\alpha_k\right) - \sum_{k=1}^{K}\log\Gamma\left(\alpha_k\right)\right) = D(K-1)\log\alpha + \mathcal{O}(1). \quad (4.15)$$

Proof. $\alpha \leq 1.0$ のとき，$\frac{1}{\Gamma(\alpha)} = \alpha + \mathcal{O}(\alpha^2)$ (Steck and Jaakkola, 2002) と近似でき，以下を得る．

$$\begin{aligned} & D\left(\log\Gamma(\sum_{k=1}^{K}\alpha_k) - \sum_{k=1}^{K}\log\Gamma(\alpha_k)\right) \\ & = D\left(\sum_{k=1}^{K}\log\alpha_k - \log\sum_{k=1}^{K}\alpha_k\right) + \mathcal{O}\left(max\left(\frac{\alpha}{KD}\right)^2\right) \end{aligned} \quad (4.16)$$

対数関数の性質より $\alpha \in (0,1)$ におけるイェンゼンの不等式は

$$\frac{1}{K}\sum_{k=1}^{K}\log\alpha_k + \log K \leq \log\sum_{k=1}^{K}\alpha_k$$

トピック数 $K \geq 1$ より，$\frac{1}{K}\sum_{k=1}^{K}\log\alpha_k \leq \log\sum_{k=1}^{K}\alpha_k$

結果として以下を得る

$$\sum_{d=1}^{D}\left(\log\Gamma(\sum_{k=1}^{K}\alpha_k) - \sum_{k=1}^{K}\log\Gamma(\alpha_k)\right) \leq \sum_{d=1}^{D}\frac{K-1}{K}\sum_{k=1}^{K}\log\alpha_k + \mathcal{O}(1)$$

$\sum_{k=1}^{K}\alpha_k = \alpha$ より

$= D(K-1)\log\alpha + \mathcal{O}(1).$

□

同様にして，$\log p(W, Z|\beta)$ の事前分布の項も漸近展開でき，以下を得る．

$$K\left(\log\Gamma\left(\sum_{v=1}^{V}\beta_v\right)-\sum_{k=1}^{K}\log\Gamma\left(\beta_v\right)\right)= \quad K(V-1)\log\beta+\mathcal{O}(1) \tag{4.17}$$

定理 1 から，ハイパーパラメータ α が 1 よりも小さいとき，$D(K-1)\log\alpha$ が事前分布項の中で支配的となる．α が 0 に近づくにつれ，事前分布の項の値は小さくなるので，トピック z_i が出現し難くなる．ハイパーパラメータ β が 1 よりも小さいとき，$K(V-1)\log\beta$ が事前分布項の中で支配的となる．β が 0 に近づくにつれ，事前分布の項は小さくなり，語彙 v にトピック z が割り当てられにくくなる．

定理 2. $\alpha \geq 1.0$ のとき，$p(Z \mid \alpha)$ の事前分布の項は，以下のように漸近展開できる．

$$D\left(\log\Gamma(\sum_{k=1}^{K}\alpha_k)-\sum_{k=1}^{K}\log\Gamma(\alpha_k)\right)=\alpha D\log K+\frac{D(K-1)}{2}\log\frac{\alpha}{2\pi K}+\mathcal{O}(1) \tag{4.18}$$

Proof. $\alpha \geq 1.0$ のとき，以下のスターリン展開を用いる．

$$\log\Gamma\left(\alpha\right)=\frac{1}{2}\log\left(2\pi\right)+\left(\alpha-\frac{1}{2}\right)\log\alpha-\alpha+\mathcal{O}\left(\frac{1}{\alpha}\right),$$

したがって，$\alpha \geq 1.0$ のとき，以下を得る．

$$\begin{aligned}D\left(\log\Gamma(\sum_{k=1}^{K}\alpha_k)-\sum_{k=1}^{K}\log\Gamma(\alpha_k)\right)=&-D\sum_{k=1}^{K}\alpha_k\log\frac{\sum_{k=1}^{K}\alpha_k}{\alpha_k}\\&-\frac{D}{2}\left((K-1)\log(2\pi)-\sum_{k=1}^{K}\log\alpha_k+\log\sum_{k=1}^{K}\alpha_k\right)\\&+\mathcal{O}(max(\frac{1}{\alpha})),\end{aligned}$$

$\alpha \geq 1.0$ より，イェンゼンの不等式は，

$$\frac{1}{K}\sum_{k=1}^{K}\log\alpha_k\geq\log\sum_{k=1}^{K}\alpha_k,$$

これより，以下を得る．

$$\sum_{d=1}^{D}\left(\log\Gamma(\sum_{k=1}^{K}\alpha_k)-\sum_{k=1}^{K}\log\Gamma(\alpha_k)\right)\geq D\alpha\log K$$
$$-\frac{D}{2}\left((K-1)\log 2\pi-\frac{K-1}{K}\sum_{k=1}^{K}\log\alpha_k\right)$$
$$=\alpha D\log K+\frac{D(K-1)}{2}\log\frac{\alpha}{2\pi K}+\mathcal{O}(1)$$

□

ハイパーパラメータ α の値が 1 より大きいとき，$\frac{D(K-1)}{2}\log\frac{\alpha}{2\pi K}$ が事前分布の項の中で支配的であり，α が増大するにつれ事前分布の項は単調増加する．結果として，α が増大するにつれ，トピックが出現しやすくなる．同様にして，$\log p(W,Z|\beta)$ の事前分布の項も漸近展開でき，以下を得る．

$$K\left(\log\Gamma\left(\sum_{v=1}^{V}\beta_v\right)-\sum_{v=1}^{V}\log\Gamma\left(\beta_v\right)\right)=\beta K\log V+\frac{K(V-1)}{2}\log\frac{\beta}{2\pi V}+\mathcal{O}(1).$$

ハイパーパラメータ β の値が 1 より大きいとき，$\frac{K(V-1)}{2}\log\frac{\beta}{2\pi V}$ が事前分布の項の中で支配的であり，β が増大にするにつれ事前分布の項は単調増加する．結果として，β が増大するにつれ，語彙 v にトピックが割り当てられやすくなりトピック数が大きくなる．

4.4.2 尤度項の分析

定理 3. $\alpha+N$ が十分大きいとき，$p(Z\mid\alpha)$ の尤度項は，以下のように漸近展開できる．

$$\sum_{d=1}^{D}\left(\sum_{k=1}^{K}\log\Gamma\left(N_{kd}+\alpha_k\right)-\log\Gamma\left(N_d+\sum_{k=1}^{K}\alpha_k\right)\right)$$
$$=\sum_{d=1}^{D}\sum_{k=1}^{K}(N_{kd}+\alpha_k)\log\left(\frac{N_{kd}+\alpha_k}{N_d+\alpha}\right)-\frac{1}{2}\sum_{d=1}^{D}\sum_{k=1}^{K}\frac{K-1}{K}\log\left(\frac{N_{kd}+\alpha_k}{2\pi}\right)+\mathcal{O}(1).$$

Proof. $\alpha+N$ が十分大きいとき，スターリンの展開式は，以下のように表せる．

$$\log\Gamma\left(\alpha\right)=\frac{1}{2}\log\left(2\pi\right)+\left(\alpha-\frac{1}{2}\right)\log\alpha-\alpha+\mathcal{O}\left(\frac{1}{\alpha}\right),$$

したがって以下を得る.

$$
\begin{aligned}
&\sum_{d=1}^{D}\left(\sum_{k=1}^{K}\log\Gamma\left(N_{kd}+\alpha_k\right)-\log\Gamma(N_d+\sum_{k=1}^{K}\alpha_k)\right)\\
=&\sum_{d=1}^{D}\left(\sum_{k=1}^{K}\left(\frac{1}{2}\log(2\pi)+\left(N_{kd}+\alpha_k-\frac{1}{2}\right)\log(N_{kd}+\alpha_k)-(N_{kd}+\alpha_k)\right)\right.\\
&\left.-\left(\frac{1}{2}\log(2\pi)+(N_d+\sum_{k=1}^{K}\alpha_k-\frac{1}{2})\log(N_d+\sum_{k=1}^{K}\alpha_k)-(N_d+\sum_{k=1}^{K}\alpha_k)\right)\right)\\
=&\sum_{d=1}^{D}\left(\sum_{k=1}^{K}(N_{kd}+\alpha_k)\log(N_{kd}+\alpha_k)+\frac{K-1}{2}\log(2\pi)\right.\\
&\left.-(N_d+\alpha_k)\log(N_d+\sum_{k=1}^{K}\alpha_k)-\frac{1}{2}\sum_{k=1}^{K}\log(N_{kd}+\alpha_k)+\frac{1}{2}\log(N_d+\sum_{k=1}^{K}\alpha_k)\right)\\
=&\sum_{d=1}^{D}\sum_{k=1}^{K}(N_{kd}+\alpha_k)\log\left(\frac{N_{kd}+\alpha_k}{N_d+\sum_{k=1}^{K}\alpha_k}\right)\\
&+\frac{1}{2}\sum_{d=1}^{D}\left((K-1)\log(2\pi)-\sum_{k=1}^{K}\log\left(N_{kd}+\alpha_k\right)+\log(N_d+\sum_{k=1}^{K}\alpha_k)\right)+\mathcal{O}\left(\frac{KD}{N+\alpha}\right)
\end{aligned}
$$

以上より, 以下を得る.

$$
\begin{aligned}
&\sum_{d=1}^{D}\left(\sum_{k=1}^{K}\log\Gamma\left(N_{kd}+\alpha_k\right)-\log\Gamma\left(N_d+\sum_{k=1}^{K}\alpha_k\right)\right)\\
=&\sum_{d=1}^{D}\sum_{k=1}^{K}(N_{kd}+\alpha_k)\log\left(\frac{N_{kd}+\alpha_k}{N_d+\sum_{k=1}^{K}\alpha_k}\right)\\
&-\frac{1}{2}\sum_{d=1}^{D}\left((K-1)\log(2\pi)\sum_{k=1}^{K}\log(N_{kd}+\alpha_k)-\log(N_d+\sum_{k=1}^{K}\alpha_k)\right)+\mathcal{O}\left(\frac{DK}{N+\alpha}\right)
\end{aligned}
$$

ここでイェンゼンの以下の不等式を用いる.

$$
\frac{1}{K}\sum_{k=1}^{K}\log(N_{kd}+\alpha_k)+\log K\geq\log(N_d+\sum_{k=1}^{K}\alpha_k)
$$

これより, $\sum_{d=1}^{D}\left(\sum_{k=1}^{K}\log(N_{kd}+\alpha_k)-\log(N_d+\sum_{k=1}^{K}\alpha_k)\right)$ の中で $\sum_{k=1}^{K}\log(N_{kd}+\alpha_k)$ が支配的になることが推測できる.

$\log(N_d + \sum_{k=1}^{K} \alpha_k)$ を上限 $\frac{1}{K}\sum_{k=1}^{K}\log(N_{kd}+\alpha_k) + \log K$ により近似し、以下を得る.

$$\sum_{d=1}^{D}\left(\sum_{k=1}^{K}\log\Gamma\left(N_{kd}+\alpha_k\right) - \log\Gamma\left(N_d + \sum_{k=1}^{K}\alpha_k\right)\right)$$
$$= \sum_{d=1}^{D}\sum_{k=1}^{K}(N_{kd}+\alpha_k)\log\left(\frac{N_{kd}+\alpha_k}{N_d+\alpha}\right) - \frac{1}{2}\sum_{d=1}^{D}\sum_{k=1}^{K}\frac{K-1}{K}\log\left(\frac{N_{kd}+\alpha_k}{2\pi}\right) + \mathcal{O}(1).$$

□

同様にして，$\log p(W|Z,K,\beta)$ の尤度項も漸近展開でき，以下を得る.

$$\sum_{v=1}^{V}\left(\sum_{v=1}^{V}\log\Gamma\left(N_{kv}+\beta_v\right) - \log\Gamma\left(N_k + \sum_{v=1}^{V}\beta_v\right)\right)$$
$$= \sum_{k=1}^{K}\sum_{v=1}^{V}(N_{kv}+\beta_v)\log\left(\frac{N_{kv}+\beta_v}{N_k+\beta}\right) - \frac{1}{2}\sum_{k=1}^{K}\sum_{v=1}^{V}\frac{V-1}{V}\log\left(\frac{N_{kv}+\beta_v}{2\pi}\right) + \mathcal{O}(1).$$

定理 3 から，尤度項は対数事後分布の項 $\sum_{d=1}^{D}\sum_{k=1}^{K}(N_{kd}+\alpha_k)\log\left(\frac{N_{kd}+\alpha_k}{N_d+\alpha}\right)$ とペナルティー項 $\frac{1}{2}\sum_{d=1}^{D}\sum_{k=1}^{K}\frac{K-1}{K}\log\left(\frac{N_{kd}+\alpha_k}{2\pi}\right)$ にわけられる.

α が十分に大きくなるとき，対数事後分布の項は α が大きくなるにつれトピックが多く出現するような働きをする．またペナルティー項も α が大きくなるにつれ増大するが，$\sum_{d=1}^{D}\sum_{k=1}^{K}\alpha_k\log\left(\frac{N_d+\alpha_k}{N_d+\alpha}\right)$ の影響の方が大きいため，結果として α が大きくなるときトピック数は大きくなる.

β が十分に大きくなるとき，対数事後分布の項の中で β の影響が大きくなり，対数事後分布の項が語彙 v がトピック z_i に割り当てられやすくなるように働く．またペナルティー項の中でも β の影響が大きくなるが，語彙 v にトピックが割り当てられやすくなる．結果として β を大きくしていくと，トピック数が大きくなるように働く.

α が十分に小さくなるとき，対数事後分布の項は $\sum_{d=1}^{D}\sum_{k=1}^{K}(N_{kd}+\alpha_k)\log\left(\frac{N_{kd}+\alpha_k}{N_d+\alpha}\right) \to \sum_{d=1}^{D}\sum_{k=1}^{K}(N_{kd})\log\left(\frac{N_{kd}}{N_d}\right)$ となり，影響が小さくなる．またこのときペナルティー項の影響が小さくなるので，トピックが出現しにくくなるように働く.

β が十分に小さくなるとき，対数事後分布の項は β の影響が小さくなり，ペナルティー項の中でも β の影響が小さくなり，ペナルティーとして働かなくなる．しかし，尤度項の影響が小さくなるので，β を小さくしていくと，語彙 v にトピックが割り当てられにくくなる．

4.4.3 周辺尤度の分析

これまで事前分布の項と尤度項に分けてそれぞれ分析を行った．ここでは，ここではそれらを足し合わせた周辺尤度を分析する．

定理 4. $\alpha + N$，$\beta + N$ が十分大きく，$\alpha, \beta \leq 1.0$ のとき，

$$\sum_{Z} \log p(W, Z \mid K, \alpha, \beta) = \sum_{Z} \left(\log p(Z \mid \alpha) + \log p(W \mid Z, \beta) \right)$$

$$\log p(Z \mid \alpha) = \sum_{d=1}^{D} \sum_{k=1}^{K} (N_{kd} + \alpha_k) \log \left(\frac{N_{kd} + \alpha_k}{N_d + \alpha} \right)$$

$$-\frac{1}{2} \sum_{d=1}^{D} \sum_{k=1}^{K} \frac{K-1}{K} \log \left(\frac{N_{kd} + \alpha_k}{2\pi\alpha_k^2} \right) + \mathcal{O}(1) \tag{4.19}$$

$$\log p(W \mid Z, \beta) = \sum_{k=1}^{K} \sum_{v=1}^{V} (N_{kv} + \beta_v) \log \left(\frac{N_{kv} + \beta_v}{N_k + \beta} \right)$$

$$-\frac{1}{2} \sum_{k=1}^{K} \sum_{v=1}^{V} \frac{V-1}{V} \log \left(\frac{N_{kv} + \beta_v}{2\pi\beta_v^2} \right) + \mathcal{O}(1) \tag{4.20}$$

定理 5. $\alpha+N$，$\beta+N$ が十分大きく，$\alpha,\beta \geq 1.0$ のとき，

$$\sum_{Z} \log p(W,Z \mid \alpha,\beta) = \sum_{Z} (\log p(Z \mid \alpha) + \log p(W \mid Z,\beta) \quad)$$

$$\log p(Z \mid \alpha) = \sum_{d=1}^{D}\sum_{k=1}^{K}(N_{kd}+\alpha_k)\log\left(\frac{N_{kd}+\alpha_k}{N_d+\alpha}\right) + \alpha D \log K$$
$$-\frac{1}{2}\sum_{d=1}^{D}\sum_{k=1}^{K}\frac{K-1}{K}\log\left(1+\frac{N_{kd}}{\alpha_k}\right)+\mathcal{O}(1) \quad (4.21)$$

$$\log p(W \mid Z,\beta) = \sum_{k=1}^{K}\sum_{v=1}^{V}(N_{kv}+\beta_v)\log\left(\frac{N_{kv}+\beta_v}{N_k+\beta}\right) + \beta K \log V$$
$$-\frac{1}{2}\sum_{k=1}^{K}\sum_{v=1}^{V}\frac{V-1}{V}\log\left(1+\frac{N_{kv}}{\beta_v}\right)+\mathcal{O}(1) \quad (4.22)$$

$\alpha \to 0$ のとき，式 (4.19) の $\sum_{d=1}^{D}\sum_{k=1}^{K}(N_{kd}+\alpha_k)\log\left(\frac{N_{kd}+\alpha_k}{N_d+\alpha}\right) \to \sum_{d=1}^{D}\sum_{k=1}^{K}N_{kd}\log\left(\frac{N_{kd}}{N_d}\right)$ となる．$N_{kd} > N_d$ より，$\log p(Z \mid \alpha)$ は小さくなり，トピックが出現しにくくなる．$\beta \to 0$ のときも同様にして，$\sum_{d=1}^{D}\sum_{k=1}^{K}N_{kv}\log\left(\frac{N_{kv}}{N_k}\right)$ が小さくなるので，語彙 v にトピックが割り当てられにくくなり，結果としてトピック数は小さくなる．ただし，$V > N_d$ のとき，$\sum_{k=1}^{K}\sum_{v=1}^{V}N_{kv}\log\left(\frac{N_{kd}}{N_d}\right)$ の影響が小さくなる．また，語彙数が少なく文章内の単語が多いためひとつのトピックに振られる語彙の数 N_{kv} が大きくなるため，そのためトピック数を大きくするように働く．

α が十分に大きくなるとき，式 (4.21) のペナルティー項の影響が減少する．しかし，$\sum_{d=1}^{D}\sum_{k=1}^{K}(N_{kd}+\alpha_k)\log\left(\frac{N_{kd}+\alpha_k}{N_d+\alpha}\right)$ が大きくなり，トピックを出現させるように働く．$\alpha D \log K$ は単調増加をしトピックを多く出現させようとする働きをする．結果として $\sum_{d=1}^{D}\sum_{k=1}^{K}(N_{kd}+\alpha_k)\log\left(\frac{N_{kd}+\alpha_k}{N_d+\alpha}\right)$ の影響が強くトピックを出現させるように働き，結果としてトピック数が大きくなる．β が十分に大きくなるときも同様にして，$\sum_{k=1}^{K}\sum_{v=1}^{V}(N_{kv}+\beta_v)\log\left(\frac{N_{kv}+\beta_v}{N_k+\beta}\right)$ の影響が強く結果としてトピック数は大きくなる．ただし，$V > N_d$ となるとき，N_k，N_k，N_{kv} が小さくなり，$\sum_{d=1}^{D}\sum_{k=1}^{K}(N_{kd}+\alpha_k)\log\left(\frac{N_{kd}+\alpha_k}{N_d+\alpha}\right)$ が小さくなりトピック数を小さくするように働く．$\sum_{k=1}^{K}\sum_{v=1}^{V}(N_{kv}+\beta_v)\log\left(\frac{N_{kv}+\beta_v}{N_k+\beta}\right)$ が語彙にトピックを割り当てない様に働くため，結果としてトピック数を小さ

くするように働く．

α が十分に大きくなるとき，式（4.21）のペナルティー項の影響が減少する．しかし，$\sum_{d=1}^{D}\sum_{k=1}^{K}(N_{kd}+\alpha_k)\log\left(\frac{N_{kd}+\alpha_k}{N_d+\alpha}\right)$ が大きくなり，トピックを出現させないようにペナルティー項として働く．$\alpha D\log K$ は単調増加をしトピックを多く出現させようとする働きをする．$\beta \to 0$ のとき，$\log p(W \mid Z, \beta)$ は小さくなり語彙 v にトピックが割り当てられにくくなり，結果としてトピック数は小さくなる．

α が十分小さいとき, 式 (4.19) の $\sum_{d=1}^{D}\sum_{k=1}^{K}(N_{kd}+\alpha_k)\log\left(\frac{N_{kd}+\alpha_k}{N_d+\alpha}\right) \to \sum_{d=1}^{D}\sum_{k=1}^{K}N_{kd}\log\left(\frac{N_{kd}}{N_d}\right)$ となる．$N_{kd} > N_d$ より，$\log p(Z \mid \alpha)$ は小さくなり，トピックが出現しにくくなる．β が十分大きいとき，$\beta K\log V$ は単調増加し，$\sum_{k=1}^{K}\sum_{v=1}^{V}(N_{kv}+\beta_v)\log\left(\frac{N_{kv}+\beta_v}{N_k+\beta}\right)$ の影響が強く結果としてトピック数は大きくなる．

また，対数周辺尤度は対数事後分布の項とペナルティー項のトレードオフと考えられ，1 を境にその働きが入れ替わることがわかる．$\alpha \leq 1$ のとき，ペナルティー項は $\frac{1}{2}\sum_{d=1}^{D}\sum_{k=1}^{K}\frac{K-1}{K}\log\left(\frac{N_{kd}+\alpha_k}{2\pi\alpha_k^2}\right)$ と表されるが，α_k と α_k^2 の大小を考えるとペナルティー項は $\frac{1}{2}\sum_{d=1}^{D}\sum_{k=1}^{K}\frac{K-1}{K}\log\left(\frac{N_{kd}}{2\pi\alpha_k^2}\right)$ と近似できる．$\alpha \geq 1$ のとき，ペナルティー項は $\frac{1}{2}\sum_{d=1}^{D}\sum_{k=1}^{K}\frac{K-1}{K}\log\left(1+\frac{N_{kd}}{\alpha_k}\right)$ である．ペナルティ項はデータとハイパーパラメータとの比になっている．この形から，α が小さくなるときトピック数は小さくなり，α が大きくなるときトピック数の推定値が大きくなると考えられる．トピック数の推定値がハイパーパラメータに敏感になる原因である．データが十分にあるとき，学習への影響を最大にするためには，ペナルティ項の中でデータを最も反映する形にすれば良い．α が 1 のとき，ペナルティ項はデータの影響を最大化でき，学習への影響を最大にできるので，データが十分にあるときハイパーパラメータは 1 が最も良い．

以上より，データが十分にあるとき，α，β の値によるトピック数の推定値は，以下のようになる．

- $\alpha = \beta = 1$ のとき，真値に近づく．
- α，β が小さいとき，過小評価される．
- α，β が大きいとき，過大評価される．
- $\alpha = 1$，β が小さいとき，過小評価される．しかし，データ大きくなりすぎると真値よりも大きくなる．
- $\alpha = 1$，β が大きいとき，過小評価される．しかし，データが大きくなりすぎると真値よりも大きくなる．
- $\beta = 1$，α が小さいとき，過小評価される．しかし，データが大きくなりすぎると過大評価となる．
- $\beta = 1$，α が大きいとき，過小評価される．しかし，データが大きくなりすぎると過大評価となる．
- α 大きく，β が小さいとき，過小評価される．しかし，データ大きくなるにつれて過大評価となる．
- α 小さく，β が大きいとき，データが大きくなるにつれて真値に近づくが α に対してデータが大きくなりすぎると過大評価となる．

4.5 レポートデータへの適用

前節までに周辺尤度最大化によるトピック数の推定値は，ハイパーパラメータ α, β に非常に敏感であり，データが十分に大きいとき $\alpha = \beta = 1.0$ とすることで高精度にトピック数を推定出来ることを示した．この結果を用いて，第 3 章で用いたデータをレポートデータのトピック数を推定することを考える．レポートデータは，文書数は 90，語彙数は 5492，各文書の単語数の平均は 312.2 である．レポートデータは，文書数が少なく，各文書の単語数よりも語彙数が非常に大きいためデータがスパースであるため，トピック数推定が困難である．このような場合の対処法として十分なデータを用意することであるが，現実的に困難である．そのため，本研究では擬似的にデータを増やすことを考える．LDA の周辺尤度の式（4.13），式（4.14）から，α_k, β_v はそれぞれ

データ N_{kd}, N_{kv} を補正する擬似データと考えられる．そのため，スパースなデータのときハイパーパラメータを大きくしデータを補正し擬似的にデータを増やすことで，トピック数を推定できると考えられる．このことをシミュレーションデータを用いて確認する．レポートデータと同様の条件のシミュレーションデータを生成し，トピック数を推定した．表 4.17 はラプラス近似による周辺尤度（式（4.10)），表 4.18 は調和平均による周辺尤度（式（4.5)）を用いてトピック数を推定した結果を示す．表 4.17 から，ラプラス近似による周辺尤度を用いたときトピック数を推定することができなかった．表 4.18 から，調和平均を用いる場合，$\alpha = 1, \beta = 10000$ のとき，真のトピック数に近い値を推定している．

表 4.17　ラプラス近似，$K = 10, D = 100, V = 5000, N_d = 300$

α	β	mse	ave	logML
0.0001	0.0001	64	2	-449491.13
0.0001	1	64	2	-311628
0.0001	10000	64	2	-173814.04
1	0.0001	64	2	-182209.06
1	1	64	2	-180816.73
1	10000	64	2	-182301.68
10000	0.0001	64	2	-8539355.78
10000	1	64	2	-8401968.11
10000	10000	62.13	5.88	-8319741.74

この結果から，調和平均を用いた周辺尤度 (式（4.5）) を用いてレポートデータのトピック数を推定した．推定結果を図 4.5.1 に示す．縦軸は logML 値を示し，横軸をトピック数とした．

表 4.18　調和平均，$K = 10, D = 100, V = 5000, N_d = 300$

α	β	mse	ave	logML
0.0001	0.0001	400	30	-213497.9
0.0001	1	400	30	-189178.82
0.0001	10000	13.4	6.6	-244429.98
1	0.0001	400	30	-212846.65
1	1	566.67	33.33	-179155.9
1	10000	5.6	9.6	-243781.01
10000	0.0001	400	30	-212583.61
10000	1	400	30	-174454.1
10000	10000	67.4	1.8	-246837.22

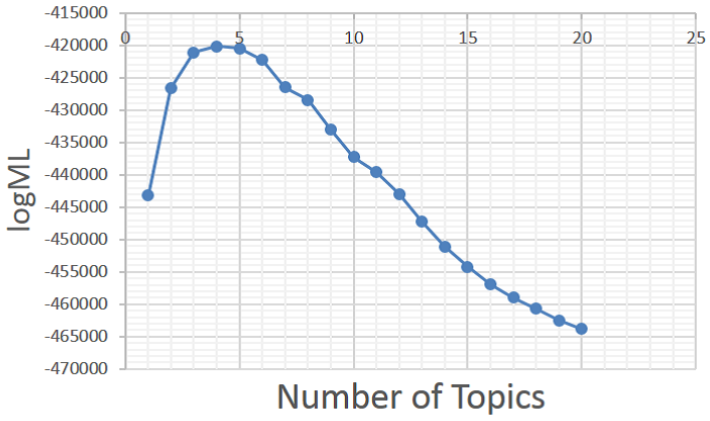

図 4.5.1　レポートデータのトピック数推定結果（$\alpha = 1, \beta = 10000$）

図 4.5.1 から，レポートデータのトピック数は 4 と推定でき，人手による分類と同じ結果となった．これにより，データが十分に大きいときは，ラプラス近似による周辺尤度によりトピック数を推定し，データがスパースな場合は，調和平均による周辺尤度を用い，ハイパーパラメータを大きく与えることで（今回のデータでは $\alpha = 1, \beta = 10000$）とすれば，トピック数を推定できる．

4.6 むすび

本章では，LDA のトピック数の推定値とハイパーパラメータの関係性について議論し，シミュレーション及び漸近解析により，トピック数の推定値がハイパーパラメータの値に敏感であることを示した．具体的には，ハイパーパラメータを小さくするときトピック数が過小評価され，大きくするときトピック数が過大評価さる．また，ハイパーパラメータが 1 としたとき，真のトピック数を推定できることを示した．レポートデータのようなスパースなデータに対しては，ハイパーパラメータを大きく与えることで（今回のデータでは $\alpha = 1, \beta = 10000$），トピック数を推定できる．その結果，人手による分類結果と同一のトピック数となった．

第 5 章

結言

本論文では，レポートライティングにおける他者からの学びを支援するために，過去の学ぶべきレポートを学習者に推薦するシステムを提案した．

第 2 章では，関連研究を紹介をした．具体的には，本システムに用いたレポートデータを蓄積している LMS（Learning Management System）“Samurai”，導入，背景，目的，方法，結論」といった形式的な構成を解析し，学習者の論文構成を可視化や指摘するシステムが主である従来のレポートライティング支援システム，学習者データと類似性が高いコンテンツや人，メッセージを推薦することが主である教育分野における推薦システムを紹介した．

第 3 章では，レポートの主題を自動的に推定できる LDA（Latent Dirichlet Allocation）を用いたレポート推薦システムを提案し，その評価について述べた．その特徴は，(1) LDA により，学習者のレポートの潜在的なトピックを推定し，他者レポートとのトピック分布の距離を計算して，同一の主題を扱う他者レポートを検索する手法を提案した．さらに，(2) 学習者のレポートと他者レポートとの単語分布の距離を計算し，同一の主題を扱うが，内容（用いられる単語分布）の異なる評価の高い他者のレポートを多様に推薦する手法を提案したことである．

本システムの有効性を示すため，実際の理工系大学生を対象に評価実験を行った．その結果，提案手法を用いると簡単には習得できないスキル，レポー

トの構成，表現，オリジナリティの改善が見られた．また，アンケートにより，提案手法の有効性を示した．

第 4 章では，LDA のトピック数の推定について述べた．第 3 章において，トピック数を専門家による評価データを用いた分類精度から決定した．しかし，データが大量になった場合や新たにデータを追加する際に人手による分類を作成しなおす必要があり，システムを利用する上で現実的ではない．また，人手による分類に即したトピック数が，モデルの学習・推定精度を高くする保証はない．そこで，本章では，トピック数を変え，LDA の周辺尤度を計算し，周辺尤度の値が最も高くなるときのトピック数をモデルの真のトピック数として決定した．周辺尤度からトピック数を決定する際，LDA のハイパーパラメータが結果に大きく影響することをシミュレーションにより示した．具体的には，ハイパーパラメータを小さくするときトピック数は過小評価され，ハイパーパラメータを大きくするときトピック数は過大評価される．このような現象が起きるメカニズムを LDA の周辺尤度を漸近解析することにより明らかにした．データが十分に大きいとき，ハイパーパラメータが 1 としたとき，トピック数を最も正確に推定できることを漸近解析及びシミュレーションにより示した．レポートデータのようなスパースなデータにおいては，ハイパーパラメータの値を大きく与えることで（今回のデータでは $\alpha = 1, \beta = 10000$），トピック数を自動的に決定できることを示した．

参考文献

[1] M.Ueno. Data mining and text mining technologies for collaborative learning in an ilms "samurai". In *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, ICALT '04, pp. 1052–1053, Washington, DC, USA, 2004. IEEE Computer Society.

[2] M.Ueno. On-line contents analysis system for e-learning. In *Advanced Learning Technologies, 2004. Proceedings. IEEE International Conference on*, pp. 762–764, Aug 2004.

[3] M.Ueno. Animated pedagogical agent based on decision tree for e-learning. In *Proceedings of the Fifth IEEE International Conference on Advanced Learning Technologies*, ICALT '05, pp. 188–192, Washington, DC, USA, 2005. IEEE Computer Society.

[4] 植野真臣, 宇都雅輝. 他者からの学びを誘発する e ポートフォリオ (< 特集 >新時代の学習評価). 日本教育工学会論文誌, Vol. 35, No. 3, pp. 169–182, dec 2011.

[5] 植野真臣. 多機能型 e ポートフォリオシステム "samurai-folio" の開発. 日本教育工学会研究報告集, Vol. 2010, No. 3, pp. 33–40, jul 2010.

[6] O'Rourke Stephen T. and Calvo Rafael A. Analysing semantic flow in academic writing. In *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*, pp. 173–180, Amsterdam, The Netherlands, The Netherlands, 2009. IOS Press.

[7] 西村健士, 島津秀雄. 特定表現の重点的解析による科学技術論文構造化手法. 情報処理学会研究報告情報学基礎（FI）, Vol. 1993, No. 39, pp. 35–42, may 1993.

[8] 甲斐郷子, 中村順一, 吉田將. 表層表現に基づく文章構造解析を利用した論文改訂支援システムの試作と評価. 情報処理学会研究報告. 自然言語処理研究会報告, Vol. 106, pp. 79–84, mar 1995.

[9] 岩田芳明, 山村毅, 大西昇. マークアップ方式による文章作成システム. 電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解, Vol. 97, No. 595, pp. 31–38, mar 1998.

[10] 山崎通弘, 山村毅, 大西昇. 選択可能なスタイルを用いた文書作成支援システム. 全国大会講演論文集, Vol. 57, pp. 211–212, oct 1998.

[11] 舘野泰一, 大浦弘樹, 望月俊男, 西森年寿, 山内祐平, 中原淳. アカデミック・ライティングを支援する ict を活用した協同推敲の実践と評価 (教育実践研究論文). 日本教育工学会論文誌, Vol. 34, No. 4, pp. 417–428, mar 2011.

[12] Chris Reed and Glenn Rowe. Araucaria: Software for argument analysis, diagramming and representation. *International Journal of AI Tools*, Vol. 14, pp. 961–980, 2004.

[13] 宇都雅輝, 植野真臣. ベイズ符号を用いた論文構成構築支援システム (教育工学). 電子情報通信学会論文誌. D, 情報・システム, Vol. 94, No. 12, pp. 2069–2081, dec 2011.

[14] 宇都雅輝, 鈴木宏昭, 植野真臣. Toulmin モデルのベイジアンネットワーク表現を用いた論証推敲支援システム (教育工学). 電子情報通信学会論文誌. D, 情報・システム, Vol. 96, No. 4, pp. 998–1011, apr 2013.

[15] K. Verbert, N. Manouselis, X. Ochoa, M. Wolpers, H. Drachsler, I. Bosnic, and E. Duval. Context-aware recommender systems for learning: A survey and future challenges. *Learning Technologies, IEEE Transactions on*, Vol. 5, No. 4, pp. 318–335, Oct 2012.

[16] Kurt D. Bollacker, Steve Lawrence, and C. Lee Giles. A system for automatic personalized tracking of scientific literature on the web. In *Proceedings of the Fourth ACM Conference on Digital Libraries*, DL '99, pp. 105–113, New York, NY, USA, 1999. ACM.

[17] Allison Woodruff, Rich Gossweiler, James Pitkow, Ed H. Chi, H. Chi, and Stuart K. Card. Enhancing a digital book with a reading recommender, 2000.

[18] Sean M. McNee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, and John Riedl. On the recommending of citations for research papers. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work*, CSCW '02, pp. 116–125, New York, NY, USA, 2002. ACM.

[19] G. Salton and C. S. Yang. On the specification of term values in automatic indexing. *Journal of Documentation.*, Vol. 29, No. 4, pp. 351–372, 1973.

[20] Tiffany Tang and Gordon McCalla. Smart recommendation for an evolving e-learning system: Architecture and experiment. *International Journal on E-Learning*, Vol. 4, No. 1, pp. 105–129, 2005.

[21] K.I. Bin Ghauth and N.A. Abdullah. Building an e-learning recommender system using vector space model and good learners average rating. In *ICALT*, pp. 194–196, 2009.

[22] Jie Lu. Personalized e-learning material recommender system. In *In: Proc. of the Int. Conf. on Information Technology for Application*, pp. 374–379, 2004.

[23] Mohamed Koutheair Khribi, Mohamed Jemni, and Olfa Nasraoui. Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval. In *ICALT*, pp.

241–245. IEEE, 2008.

[24] Fabian Abel, Ig Ibert Bittencourt, Evandro de Barros Costa, Nicola Henze, Daniel Krause, and Julita Vassileva. Recommendations in online discussion forums for e-learning systems. *IEEE Transactions on Learning Technologies (TLT)*, Vol. 3, No. 2, pp. 165–176, 2010.

[25] Jie-Chi Yang, Yi-Ting Huang, Chi-Cheng Tsai, Ching-I Chung, and Yu-Chieh Wu. An automatic multimedia content summarization system for video recommendation. *Educational Technology and Society*, Vol. 12, No. 1, pp. 49–61, 2009.

[26] Yueh-Min Huang, Tien-Chi Huang, Kun-Te Wang, and Wu-Yuin Hwang. A markov-based recommendation model for exploring the transfer of learning on the web. *Educational Technology and Society*, Vol. 12, No. 2, pp. 144–162, 2009.

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

[28] 植野真臣. 過去の学習者履歴データを利用した e ポートフォリオ・システム. 情報知識学会誌, Vol. 24, No. 4, pp. 414–423, 2014.

[29] 鈴木宏昭, 杉谷裕美子. レポートライティング教育の意義と課題, 学びあいが生み出す書く力：大学におけるレポートライティング教育の試み. 丸善プラネット（株）, 2009.

[30] L.S. Vygotsky and M. Cole. *MIND IN SOCIETY*. Harvard University Press, 1978.

[31] 植野真臣. 他者からの学びの支援（特集「学習科学と学習工学のフロンティア—私の“学習”研究—（後編)」にあたって）. 人工知能学会論文誌, Vol. 30, No. 4, pp. 469–472, july 2015.

[32] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMA-*

TION SCIENCE, Vol. 41, No. 6, pp. 391–407, 1990.

[33] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pp. 50–57, New York, NY, USA, 1999. ACM.

[34] 椿本弥生, 赤堀侃司. 主観的レポート評価の系列効果を軽減するツールの開発と評価. 日本教育工学会論文誌, Vol. 30, No. 4, pp. 275–282, mar 2007.

[35] 椿本弥生, 柳沢昌義, 赤堀侃司. レポート内容とその評価を可視化する円錐形レポート採点支援マップの開発と評価 (<特集 >学習オブジェクト・学習データの活用と集約). 日本教育工学会論文誌, Vol. 31, No. 3, pp. 317–326, dec 2007.

[36] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, Vol. 101, No. Suppl. 1, pp. 5228–5235, April 2004.

[37] T. Minka. Estimating a dirichlet distribution. Technical report, MIT, 2000.

[38] M. Welling A. Asuncion, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. *In UAIf09: Proceedings of the International Conference onUncertainty in Artifficial Intelligence, 2009.*, 2009.

[39] 工藤拓, 山本薫, 松本裕治. Conditional random fields を用いた日本語形態素解析 (解析), may 2004.

[40] David Arthur. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pp. 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[41] Matt Taddy. On estimation and selection for topic models. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence*

and Statistics, AISTATS 2012, La Palma, Canary Islands, April 21-23, 2012, pp. 1184–1193, 2012.

[42] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, Vol. 101, , 2004.

[43] Hanna M. Wallach. *Structured Topic Models for Language*. PhD thesis, University of Cambridge, 2008.

[44] Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pp. 1105–1112, New York, NY, USA, 2009. ACM.

[45] Wray Buntine. Estimating likelihoods for topic models. In *Proceedings of the 1st Asian Conference on Machine Learning: Advances in Machine Learning*, ACML '09, pp. 51–64, Berlin, Heidelberg, 2009. Springer-Verlag.

[46] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the american statistical association*, Vol. 90, No. 430, pp. 773–795, 1995.

[47] David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. Distributed algorithms for topic models. *J. Mach. Learn. Res.*, Vol. 10, pp. 1801–1828, December 2009.

[48] M.Ueno. Learning networks determined by the ratio of prior and data. In *UAI*, 2010.

[49] M.Ueno. Robust learning bayesian networks for prior belief. In *UAI*, 2011.

謝辞

本研究を進めるにあたり, 終始懇切なる御指導を賜った, 電気通信大学大学院教授の植野真臣先生に, 心より感謝を申し上げます. 本論文の審査過程において, 数々の貴重な御助言と御指導を賜りました栗原聡教授，大須賀昭彦教授，広田光一教授，田原康之准教授，川野秀一准教授に深謝申し上げます. また, 本研究における議論・検討に当たって、ご教示とご激励を頂いた電気通信大学大学院植野真臣研究室の西山悠助教，宇都雅輝助教，首都大学東京の石井隆稔助教，電気通信大学大学院植野研究室の皆様に心より感謝を申し上げます.

関連論文の印刷公表の方法及び時期

査読付き論文（本学位申請論文関連論文）

加藤嘉浩, 石井隆稔, 宮澤芳光, 植野真臣（2016）Latent Dirichlet Allocation を用いたレポート推薦システム，電子情報通信学会　和文 D　2016 年 2 月 Vol.J99-D,No.2,pp.152-164.

国際会議

Y. Kato, M. Ueno (2014) “E-PORTFOLIO RECOMMENDATION SYSTEM USING LDA”, 8th International Technology, Education and Development Conference (INTED) Proceedings, pp.707-716, 10-12 March 2014, Valencia, Spain.

国内学会

加藤嘉浩，石井隆稔，植野真臣（2014）LDA を用いたレポート推薦機能を持つ e ポートフォリオシステム，日本テスト学会 第 12 回全国大会論文集，pp.96-97

加藤嘉浩，石井隆稔，植野真臣（2014）LDA を用いたレポート推薦機能を持つ e ポートフォリオシステム，教育システム情報学会 第 39 回全国大会論文

集，pp.365-366

加藤嘉浩，石井隆稔，植野真臣 (2014) トピックモデルを用いたレポート推薦機能を持つ e ポートフォリオシステム，日本教育工学会　第 30 全国大会論文集，pp.389-390

杉山 剛，加藤嘉浩，石井隆稔 (2015) 適応型テストのための LDA を用いた項目間類似度の利用可能性，日本テスト学会 第 13 回全国大会論文集，pp.104-107

参考論文

Louvigne.S, Jie.Shi, Y.Kato, N.Rubens, M.Ueno, ”A corporal and LDA analysis of abstracts of academic conference papers”, Advanced Mechatronic Systems (ICAMechS), 2013 International Conference on , pp.412-416, 25-27 Sept. 2013

Louvign.S, Y.Kato, Rubens.N, Ueno.M, “Goal-Based Messages Recommendation Utilizing Latent Dirichlet Allocation”, In Advanced Learning Technologies (ICALT), 2014 IEEE 14th International Conference on, pp.464-468, 7-9 July. 2014

Louvign.S, Y.Kato, Rubens.N, Ueno.M, ” SNS Messages Recommendation for Learning Motivation”, In Artificial Intelligence in Education (AIED), pp.237-246, 22-26 June.2015, Springer International Publishing.