Vol.2022-CH-128 No.8
2022/2/19

日本中世古記録を対象としたトピック抽出自動化システムの構築

鳥居克哉[1]　中村覚[2]　山田太造[2]　稗方和夫[1]

概要：本研究では，日本史学者の史料研究支援のために，史料群に対する可用性と有用性を高めるトピック抽出を自動で行うシステムの開発を行った．ルールベースにより抽出した人名及び N-gram や Sentencepiece によって分割した用語から Bag-of-Word を生成し，LDA(Latent Dirichlet Allocation)を適用することでトピック分析を行った．さらに，史料と人物索引表を入力としてこの一連の分析を行う Web システムをクラウド上に構築した．また，鎌倉時代の公卿である藤原（勘解由小路）経光が記した『民経記』を対象にこのシステムを利用し，トピック分析の結果が史実に整合していることが確認でき，有効性が示された．

キーワード：トピックモデル，LDA，日本史，古記録，オンライン分析

Development of an Automated Topic Extraction System for Ancient Japanese Medieval Records

KATSUYA TORII[†1]　SATORU NAKAMURA[†2]
TAIZO YAMADA[†2]　KAZUO HIEKATA[†1]

Abstract: In this study，we developed a system that automatically extracts topics to increase the availability and usefulness of historical documents to support Japanese historians in their research on historical documents．We generated a Bag-of-Words from the names of people extracted by the rule base and the terms divided by N-gram and Sentencepiece.，and applied LDA (Latent Dirichlet Allocation) to analyze the topics．In addition，we constructed a web system on the cloud to perform this series of analysis using historical documents and a person index table as input．In addition，we used this system to analyze the "Minkeiki" written by Fujiwara (Kadenokoji) Tsunemitsu，a kuge of the Kamakura period，and confirmed that the results of the topic analysis were consistent with the historical facts，demonstrating its effectiveness.

Keywords: Topic Model，LDA，Japanese History，Old Diary，OLAP

1. はじめに

日本史学者は，歴史資料（以下，史料）を収集・読解・分析することで研究課題を解明していく[1]．特に分析の過程では史料を分析する際に大量の史料を読み込み，その中から研究課題に当てはまるトピックを見つける必要がある．本研究では，データ駆動型人文学研究の一環として，自然言語処理技術により日本史学者の資料研究の分析過程を支援することを目的とする．本研究では史料から抽出した人物・語句を対象としてトピック抽出を行ない，その結果をトピックに属する語句と人物のリストや時系列関係で可視化した．また，ユーザーがアクセスし一連の分析を行うことができる Web Application として開発を行なった．ケーススタディでは，鎌倉時代の公卿である藤原（勘解由小路）経光が記した『民経記』の史料を対象に，システムを用い日本史学者へのインタビューを通じてトピック抽出の評価・検討を行った．本稿は[2]をベースとして，語句抽出手法を再検討し，システムを改良した．さらに，日本史学者へのインタビューを通じてシステム評価を行った．

†1　東京大学大学院新領域創成科学研究科
Graduate School of Frontier Sciences, The University of Tokyo
†2. 東京大学史料編纂所
Historiographical Institute, The University of Tokyo

2. 関連研究と本研究の位置付け

2.1 トピックモデルのデータ分析への適用事例

本研究で扱うトピックモデルは大規模なデータを分析するための手法として広い分野で利用されているアプローチである．テキストデータだけではなく，顧客購買データ解析，画像処理，人の行動解析など様々な分野で用いられている．Sukhija ら[3]は，トピックモデルのうち，LDA(Latent Dirichlet Allocation)[4]を社会科学データに適用し，消費活動と人種・階級・ジェンダーとの関係を抽出，可視化し社会科学的な分析を行った．Heng ら[5]はアマゾンのレビューに LDA を適用し，消費者の食品購入に影響を与える要因を分析した．トピックモデルを実行する上で必要な単語の多重表現である Bag-of-Words に，どのデータを選択するかによって，多種多様な分析が可能である．

トピックモデルを歴史研究に応用した先行事例として山田ら[6]の研究が挙げられる．山田らは史料内の人物を対象に LDA を特定の史料に適用することで，史料中の人物の共起関係を基に潜在する意味関係を検出し，人物間の関係性を検出した．また，それを発展させ，時空間変化と人物の関係を抽出した．

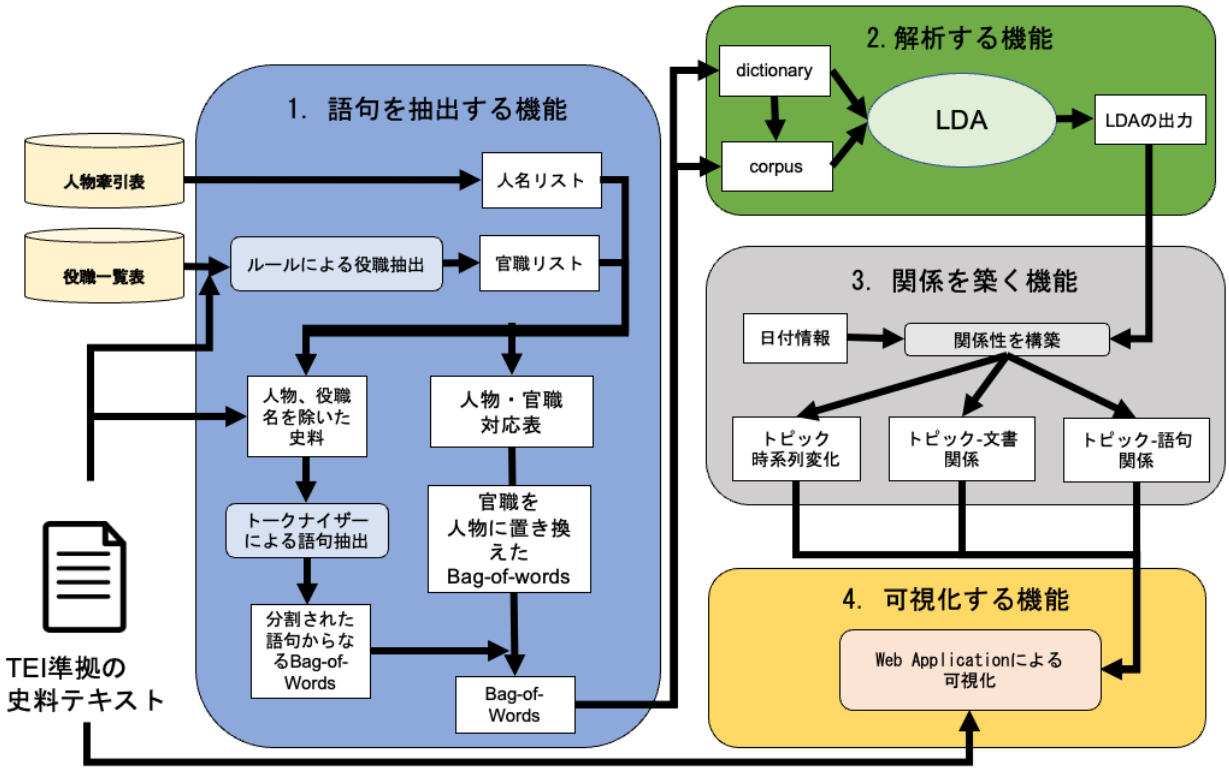

図 1　システムの概要図

Figure1　Overview of the proposed System

2.2 本研究の位置付け

先行研究では Bag-of-Words に人物だけではなく，史料に出現する語句もトピックモデルに組み込むことで，各文書のトピックを推定する．

また，既存の支援システムとして，Voyant Tools や KHcoder などがあげられるが，これらの既存のシステムとの差別化を図りながら，UI の開発を行う．

また，本研究では平安中期〜鎌倉中期の変体漢文の史料を対象にユーザーのユーズに即時対応可能な OLAP 分析システムの開発を目指す．

3. 開発したシステム

開発したシステムの概念図を図 1 に示す．ユーザーは TEI に準拠した史料と人物牽引表・役職一覧表をシステムの入力とすることで可視化された結果を取得することができる．なお，TEI に準拠した史料とは，構造化ルールを定める TEI ガイドラインに基づいてタグ付け及び構造化が行われた史料のことで，人物牽引表とは史料を編纂する際に作成される史料中に出現する人物名と出現場所を記したデータである．

以下, 開発したシステムを 4 つの機能に分けて説明する．

3.1 語句を抽出する機能

トピック抽出において史料の解析を実行する上で必要な，単語の多重表現である Bag-of-Words を作成する．まず人物牽引表から人物リストを作成し，役職一覧表を元に，史料中の役職パターンに一致する表現を抽出し，史料中の官職を人物名に置き換えた人物のみからなる Bag-of-Words を作成する．なお，本ケーススタディでは＜役職（家名+諱）＞というパターンを採用した．一方，史料から役職名を除いた史料からトークナイザーによる単語抽出を行うことで抽出した語句のみからなる Bag-of-Words を作成する．これらを組み合わせ, 人物・史料中の語句が含まれる Bag-of-Words を作成できる．

本稿では語句を抽出する際，トークナイザーとして教師なしの手法である N-gram と Sentencepiece[7]を比較検討した．

N-gram では，N=2，3 の分割で抽出された，出現回数 50 以上の出現頻度の高い語句のうち，Web 上で公開されている百科事典（コトバンク）に存在する語句を選定した．

Sentencepiece は，教師なしトークナイザーで主にニューラルモデルの学習前で語彙の大きさがあらかじめ定められている場合に用いられる手法である．テキストをサブワードに分割するため，語彙数を小さくしつつ未知語をなくすことができるという特徴がある．また，文章から直接学習

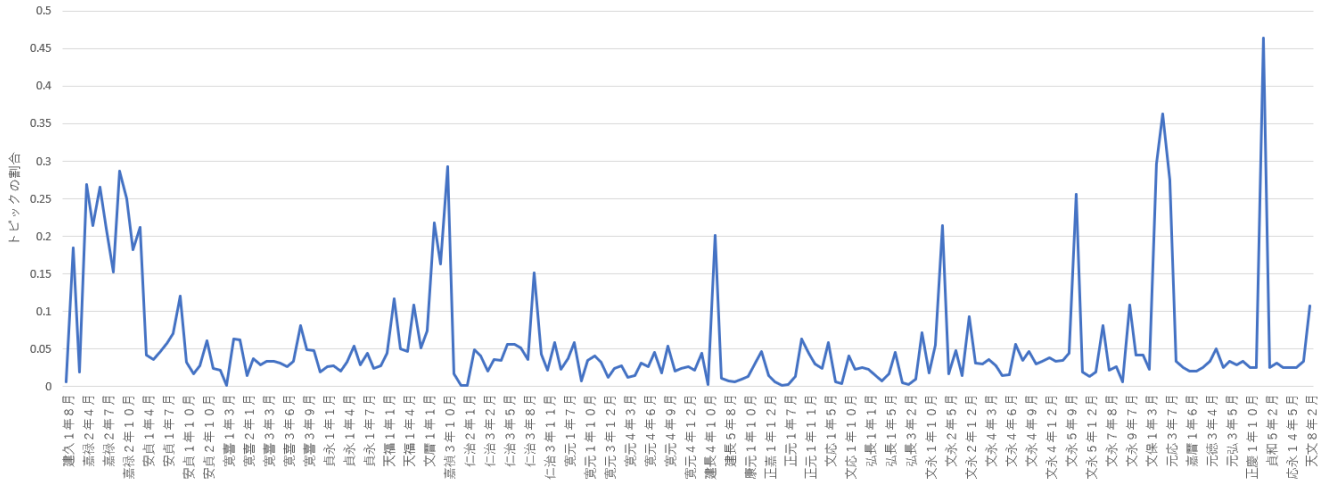

図 4　トピック 3 の月別の時系列変化
Figure4　Monthly Time Series Change Graph for Topic3

することができ，言語固有の前処理･後処理に依存しない，エンドツーエンドのシステムを構築することができる．

『民経記』の嘉禄 2 年 4 月分のテキストデータについて人手で調査した正解件数 227 件に対する結果を下記に示す．なお，N-gram においては『民経記』に対応するよう後処理を行なったが，Sentencepiece はどの史料にも対応できるように使用した分，精度が大きく落ちている．

Sentencepiece の精度が上がるように今後改善する計画ではあるが，本稿のケーススタディでは N-gram による手法を用いた．

	適合率	再現率	F値
N-gram	0.92040	0.81858	0.86651
Setencepiece	0.43096	0.40517	0.41767

図 2　語句抽出の結果
Figure2　Results of Word extraction

3.2 解析する機能

トピックモデルは対象とするデータから潜在する話題(トピック)を検出することができる教師なし学習の一つである．本研究では，トピックモデルとして，LDA を選定し，史料に出現する人物と語句を史料の特徴語として捉え，LDA の変数として組み込むことで，各史料のトピックを推定する．本研究における LDA による史料の生成確率は次式のとおりである．

$$p\left(d\middle|\alpha,\ \beta\right)$$

$$= \int Dir(\theta|\alpha)(\prod_{n=1}^{|d|}\sum_{k=1}^{C} p\left(w_n\middle|z_k,\ \beta\right)p(z_k|\theta))d\theta \quad (1)$$

αとβはパラメータ，$z = z_1,\ z_2,\ \ldots,\ z_c$は潜在トピック，$\theta = \theta_1,\ \theta_2,\ \ldots,\ \theta_c$は潜在トピックの生成確率，$Dir(\theta|\alpha)$はディレクレ分布，$d = (w_1,\ w_2 \ldots,\ w_{|d|})$は史料，$w_n$は特徴語，|d|は史料 d の単語数を表す．

LDA の計算にはオープンソースライブラリである Gensim を用いた．Gensim では，トピックモデルの推定方法のうち，従来の変分ベイズを応用したオンライン変分ベイズ[7]を使用することで，高速な学習を実現している．

入力には語彙と ID，出現回数の対応表である dictionary と dictionary に登録されている ID からなる Bag-of-Words である corpus が必要であり，3. 1 で作成した Bag-of-Words から dictionary と corpus を作成した．また，passes については，本研究では default 値として perplexity が凡そ収束する 50 を設定した．本ケーススタディにおいて Perplexity が収束する様子を図 3 に図示する．

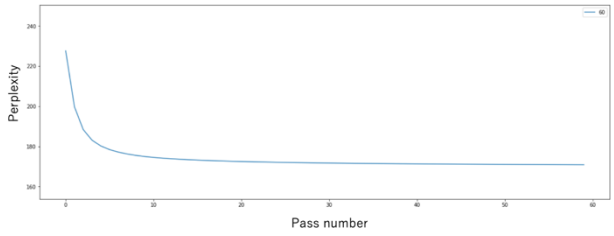

図 3　Perplexity の収束
Figure3　Perplexity changes with the number of training passes

3.3 関係性を築く機能

LDA を実行後の出力を整備し，UI に表示すべき内容を作成する．ここではトピックに対して関連度が高い単語を再び語句と人物に区別した．また，日記である対象史料から日付情報を抽出し，トピックの時間変化抽出を行った．図 4 はトピック数 10 の場合の Topic3 の月別の時系列変化グラフである．

図 5　UI の例．語句と人物のリストとトピックの時系列変化
Figure5　Monthly Time Series Change Graph for Topic3

3.4 Web で可視化する機能

トピックごとの語句，語句ごとのトピックの割合，史料ごとのトピックの割合を表示する３つの画面を作成する．データをリストやグラフの形で表示し，史料中の語句をトピックで色付けし，リンク機能を実装することでユーザビリティの向上を目指した．また，Web Application 上で入力フォームを用意し，ユーザーが入力後，同じサイト内で結果を確認することを可能とした．図５は Web Application のトピック数 10 における Topic３における UI である．画面の左側に Topic３における重要な人物と語句のリスト，右上にトピックの時系列変化，右下に関連度が高い日付の史料を可視化している．

4. ケーススタディ

4.1 『民経記』

本研究のケーススタディとして，『民経記』を対象とする史料とする．『民経記』は鎌倉中期の公家である，藤原(広橋)経光(1212~1274)の日記であり，別に『経光卿記』，『経光卿暦記』，『経光御記』，『中光記』，『民経御記』等の称呼がある．自筆原本 42 巻は，国立歴史民俗資料館蔵史料に蔵する．『大日本古記録民経記』(全十一冊)として編纂史料集が発刊されている． ケーススタディでは，編纂所がサービスしている古記録フルテキストデータベースに格納されている民経記データを使用した．文字数は 1,226,975 文字，日条数は 6630 であった．

Topic1		Topic2		Topic3		Topic4		Topic5		Topic6		Topic7		Topic8		Topic9		Topic10	
人名	語句	人名	語句	人名	語句	人名	語句	人名	語句	人名	語句	人名	語句	人名	語句	人名	語句	人名	語句
藤原経光	評定	幸清	神吉	定玄	法印	源通方	奉行	藤原経光	殿上	藤原実氏	御方	藤原実経	目録	藤原経光	沙汰	源具実	結願	藤原公相	僧正
藤原経俊	参院	藤原兼頼	木沿	平兼子	導師	藤原教実	延候	藤原隆親	此間	藤原隆親	御車	藤原師継	中原	藤原頼資	仰下	藤原盛兼	雑事	藤原伊頼	法印
平国朝	文書	惟子内親王	天恩	藤原宗氏	綱宣	藤原隆親	参内	平知宗	退出	藤原公経	供奉	藤原兼平	臨時	宣旨	謹言	平経高	雨降	荒木田永元	勘文
藤原頼資	候へ	藤原公子	天晴	相円	女房	藤原道家	相触	藤原忠高	参進	藤原道家	御所	時継	奉行	藤原家実	子細	藤原頼資	院司	宰相局	宣下
藤原経長	御所	藤原頼資	月徳	藤原頼資	其後	藤原信盛	仰下	親長	起座	藤原実有	雑色	安倍資季	賀茂	大蔵	不可	藤原光俊	奉行	藤原経光	御産
藤原光国	正月	盛家	拝官	源師親	天晴	藤原忠高	内覧	藤原良平	次第	源師房	女房	藤原親冬	後聞	源仲業	恐々	定雅	御布施	藤原隆親	文永
藤原兼経	奉行	藤原経光	歳徳	大中臣隆通	次第	平有親	行事	安倍資季	出御	賀茂在盛	装束	藤原基忠	除目	橘以良	用途	藤原定高	参行	性淵	山門
藤原定嗣	沙汰	藤原佶子	母倉	藤原宣実	布施	藤原兼高	奏聞	源顕定	其後	藤原家実	束帯	中原光朝	外記	平家守	下知	安倍季尚	行事	円実	沙汰
藤原公基	日至	藤原資定	大小	盛重	御願	家光	外記	源通忠	瓶子	中原俊職	入御	藤原良実	上皇	藤原道家	不審	平範輔	僧名	良快	衆徒
藤原顕朝	上皇	大中臣公行	土公	大中臣隆隆	僧正	藤原家嗣	沙汰	中原雅親	黄子	信光	遷御	藤原兼頼	参入	和気時定	講文	藤原実世	天晴	藤原実冬	注文
藤原高嗣	供養	藤原兼仲	裏書	藤原兼仲	退出	藤原資頼	宣命	藤原隆光猶母	吉書	平繁俊	直衣	中原資宣	院司	藤原親説	当時	家光	法勝寺	宣旨	法親王
嵯峨天皇	下官	藤原道家	加冠	藤原能教	別事	藤原親朝	此間	藤原信盛	御簾	平輔兼	騎馬	藤原経任	初度	安楽	日可	藤原公基	最勝寺	尊性法親王	摂津国
源通氏	床子	藤原実氏	以後	聖憲	装束	藤原実氏	天晴	大中臣永隆	束帯	藤原泰通	見物	賀茂在職	参仕	藤原陳子	勅定	高実	御寺	信昭	不可
安楽	延引	藤原通雅	雨下	尋	官使	平経高	退出	藤原宗平	障子	永親	松明	源資平	外宮	卜部兼直	言上	藤原為経	淵酔	藤原公親	所詮
仁明天皇	御参	藤原実基	日出	荒木田延季	金堂	平範輔	下知	藤原実世	中門	利子内親王	行幸	源雅言	宣下	藤原有長	御教書	言家	長講堂	行有	出来
紀実兼	参候	遵覚法親王	奉幣	源具守	進物	藤原光俊	宣下	藤原光業	東面	賀茂在継	相具	亀山天皇	仙洞	高階経時	領状	薬師	伝聞	源通成	勤仕
藤原兼仲	寛喜	藤原家実	日入	経円	聴衆	藤原実雄	参入	藤原実任	列立	藤原隆行	舎人	藤原基平	博士	藤原佐清	披露	藤原宣経	御祭	道融	今度
藤原宗親	自筆	円助法親王	不問	良盛	成敗	中原師兼	相催	藤原伊平	参入	藤原家嗣	布衣	藤原為氏	安倍	源有行	所労	藤原実基	殿上	藤原経範	綱所
宗尊親王	目録	暉子内親王	朝間	円憲	貞応	平範頼	日時	藤原公光	気色	藤原家季	出車	良教	職事	藤原経朝	不及	賀茂兼宣	所労	橘以隆	菅原
藤原頼長	元年	清原元久	不視病	藤原任子	停止	源頼朝	職事	源有教	西面	藤原為家	女院	藤原忠方	内宮	隆経	相尋	高経	行間	藤原行長	相論
高倉	法勝寺	薬師	禅林寺	公秉	宿所	中原親俊	数刻	橘親氏	御座	藤原長子	御輿	藤原長雅	清書	源貞元	左右	藤原隆親	免者	頼円	辞退
藤原宗雅	出御	鳥羽重久	還御	釈迦	承久	藤原宗平	御覧	源顕平	下臈	季範	出御	後嵯峨天皇	参陣	源信綱	見参	藤原実親	用途	道深法親王	御経
藤原隆親	自今	嵯峨天皇	下食	尊海	閑談	安倍国道	出納	平有親	北面	藤原陳子	頭書	藤原通雅	節会	孔子	行幸	藤原経通	仏名	兼智	興福寺
道深法親王	員数	仁明天皇	手足	良遍	浄衣	藤原有長	勘文	平時高	妻戸	源維長	陰晴	藤原隆顕	内給	藤原定嗣	来月	源定通	織物	行職	裏書
清原頼尚	御事	亀山天皇	狼藉	荒木田成定	供給	藤原為家	天皇	宗基	其儀	藤原忠高	院司	藤原公孝	叙位	藤原基通	日次	藤原基家	布施	厳縁	中臣
丹後	天福	藤原嬉子	解除	審実	建久	藤原泰通	装束	長清	上臈	大江重房	持明院	嵯峨天皇	御給	尾張	延引	源通成	先例	慶政	法眼
藤原高朝	春日	知仁親王	入来	大中臣能隆	休息	橘親氏	毎事	定宗	大進	章親	調進	源通持	仁治	藤原盛長	恐惶	藤原宣実	長吏	菅原在宗	律師
順徳天皇	不可	藤原家経	厳対	藤原実基	散花	和気時定	御所	藤原兼経	職事	源通方	中原	源具氏	臨時祭	藤原兼光	候者	藤原家良	大嘗会	藤原頼経	恒例
慈禅	不快	綾子内親王	遠行	後鳥羽天皇	御分	佐伯康長	今度	源仲業	相従	藤原基氏	行啓	源具房	如元	堀河天皇	注進	藤原兼経	修正	藤原公経	修造
四条天皇	終日	藤原経俊	結婚	盛家	鳥羽	源家清	神宝	源定通	長押	秀仁親王	召具	藤原家経	行事	幸清	大嘗会	上道時兼	除服	実性	代々

図 7　トピックに関連する語句と人物 (オレンジ：筆者関連，黄色：政務関連，青：神仏関連，ピンク：皇族関連，緑：暦注)

Figure7　Words and names of people related to the topic (orange: Author's relatives，yellow: Government affairs，blue: Religious affairs，pink: royal family，green: Calendar notes)

4.2　語句抽出評価

『民経記』における役職と語句の抽出精度について述べる．パターンマッチング+N-gram の手法で役職を抽出した場合の嘉禄 2 年 4 月分の出力件数は 53 件，人手で作成した正解件数は 64 件だった．N-gram とコトバンクによって抽出した場合の嘉禄 2 年 4 月分の出力件数は 201 件，人手で調査した正解件数は 227 件だった．それぞれの適合率，再現率，F 値を図 6 に記す．

	適合率	再現率	F 値
役職	0. 6875	0.6000	0. 6408
語句	0.9253	0.8193	0.8691

図 6　役職・語句抽出の結果

Figure6.　Results of data preprocessing

4.3　トピックの内容に関する検証

システムで出力した結果を元に，評価・検討を行った．

例としてトピック数を 10 とし，LDA の結果の検証を行う．図 7 は 10 のトピックの中から考察のために選択したトピックと，上位 10 語句と人物である．日本史学者のアドバイスを元にカテゴリ別で色分けを行った．

Topic1 は天皇関係のトピックである．父である頼資がいることからも，経光の親族と天皇間のやりとりのトピックであることがわかる．

Topic2 は暦注と経光の親族，天皇関連のトピックであることがわかる．暦注が登場することからよく登場する人名が並ぶ．Topic1 と Topic2 からは親族の情報を得ることができた．Topic3 と Topic10 は神仏関係の語句と人名が多いことから神仏のトピックであるとわかる．Topic4~Topic7 は政務関係のトピックである．これらのトピックの差は，専門家以外が Table7-6 のみで判断するのは難しい．Topic8 は書状に関するトピックであり，ここから「紙背文書」の情報が得られる．Topic9 は神仏関係のトピックとも捉えられるが，「朝儀典礼」の情報も抽出できる．「行事」「奉行」「御祭り」「大嘗祭」などの朝儀典礼関連の単語が見られるからである．朝講堂などの寺社の地名も，長講堂御八講などの行事に関係している．

また，各トピックに関して，時系列変化と対応させることで，トピックの割合が高くなる年代を特定し，史料の文章と照らし合わせ整合性を確認した．例えば，神仏関係のトピックである Topic 3 の時系列変化である 図 4 では，2 つ目の山である嘉禎元年 10 月と嘉禎 3 年 10 月に，仏教行事である「維摩会」の記述があるのが確認できた．

日本史学者へのインタビューによるシステムの検証を行った．Web Application を使ってトピックの妥当性について検証していただいた．システム全体に対してのフィードバックが得られ，分析結果が有用であり，UI がわかりやすいという評価が得られた．さらに，システムの改善すべき点として，日記数を反映させた時系列グラフの作成，紙背文書や裏書などを除外した状態で分析の必要性などがあげられた．

5. 考察

語句・役職抽出について，ある程度の精度は確認することができたが，精度向上の余地がある．今回はケーススタディにて Sentencepiece を用いることはなかったが，今後後処理に外部辞書を利用するなど，この技術を適切に用いることで精度向上を目指したい．

システム全体について，人物だけではなく語句も LDA に組み込むことにより，即座にトピックの理解ができるようになった．また，史料に含まれていることが確認されている史実を対象に，トピック分析を行った結果その史実に整合したトピックが割り当てられていることが確認され，有用性が示された．

6. おわりに

本稿では史料から N-gram を用いることでトピック抽出の対象とする人物・語句の抽出を行い，それらを Bag-of-Words としてトピック抽出を行なった．また，一連の動作を Web Application として自動で行うシステムとして開発し，『民経記』を対象に動作の検証とインタビューによる質的評価の結果から有用性を示し，資料研究の分析過程を支援することができた．

今後の研究計画としては，より柔軟性の高く，持続可能なシステムの設計及び開発を行う．具体的には，Web Application 上でユーザーが Bag-of-Words に入れる語彙の選定や，史料によって異なる役職パターンの入力を行い，その後情報学的分析結果の確認が自由にできるようなシステムの開発を行う予定である． さらに，システムを継続して利用し，恒常的に日本史学者の史料研究を支援するようなシステムの設計を行い，提案したい．

謝辞 本研究の成果の一部は JSPS 科研費 18H03576 および 19K20626 の助成を受けたものである

参考文献

[1] 中村覚:デジタルアーカイブと Linked Data を用いた歴史学研究支援に関する研究， デジタル・ヒューマニティーズ 1 29- 43 2019.
[2] 鳥居克哉，中村覚，山田太造，稗方和夫「日本史学者の要求分析に基づく歴史資料のトピック推定システムの開発」，情報処理学会第 83 回全国大会．Vol.83，No.2ZC-05，pp.1-2，2021 年.
[3] N. Sukhija, M. Tatineni, N. Brown, M. V. Moer, P. Rodriguez and S. Callicott: Topic Modeling and Visualization for Big Data in Social Sciences, 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress, Toulouse, pp. 1198-1205, 2016.
[4] D. M. Blei, A. Y. Ng, and M. I. Jordan: Latent Dirichlet Allocation, Journal of Machine Learning Research, vol. 3, pp. 993-1022(2003).
[5] YanHeng, Zhifeng Gao, Yuan Jiang, Xuqi Chen: Exploring hidden factors behind online food shopping from Amazon reviews: A topic mining approach, Journal of Retailing and Consumer Services, Volume 42, May 2018, Pages 161-168, 2018.
[6] Kudo、T.， & Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808．06226*, 2018
[7] Matthew D. Hoffman, David M. Blei, and Francis Bach. 2010. Online learning for Latent Dirichlet Allocation. In Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1 (NIPS'10). Curran Associates Inc.， Red Hook， NY， USA， 856–864.