2ZC-05

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

鳥居克哉[†1]　中村覚[†2]　山田太造[†2]　稗方和夫[†3]

東京大学工学部システム創成学科[†1]　東京大学史料編纂所[†2]

東京大学大学院新領域創成科学研究科[†3]

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]．分析の過程では、日本史学者は大量の史料を効率よく分析する必要があり，この過程を支援するシステムが必要であるのにも関わらず，未だ研究事例が少なく充分に支援されていると言えない．本稿では，トピックモデルを用いることで史料の自動分類を行い，その結果を可視化することで歴史学的見地にもとづいた史料分析を支援するシステムの提案を行う．例として，鎌倉中期の公家の広橋（藤原）経光の日記(1212~1274)である『民経記』を対象に，評価・検討を行った．

2．日本史学者の要求分析

システム方法論の手法に基づき，日本史学者の要求分析をした結果，日本史学者の要求として分析結果の有用性と理解しやすさの向上があげられることから、史料分析の支援には，史料からトピックを検出し，結果をわかりやすく提示する UI が必要であることがわかった．

3．史料へのトピックモデルの適用

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 1 つである．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[2]の研究が挙げられる．山田らは史料内の人物の関係を明確にすることはその時代の歴史像を解明することにつながると考え、史料に LDA（Latent Dirichlet Allocation）[3] を適用することで，テキスト内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．本稿では，人物だけではなく，史料に出現する語句も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．本稿における LDA による史料の生成確率は次式のとおりである．

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

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，$Dir(\theta|\alpha)$はディレクトリ分布，$d = (w_1, w_2 \ldots, w_{|d|})$は史料，$w_n$は特徴語，|d|は史料 d の特徴語数を表す．

また，史料には時間データがあることからトピックの時系列変化を抽出する．

4．　システム概要

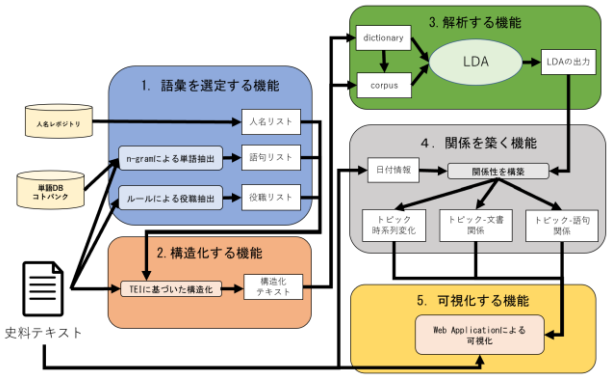

図 1：システムの概念図

開発したシステムを 5 つの機能に分けて説明する．

4.1　語彙を選定する機能

LDA を実行する上で必要な Bag-of-Words に含める語句を選定する．本稿では史料を理解する上で重要である人名と語句を選定した．また，人名の言い換え表現による重複・抜け落ちを回避するために，役職表現の抽出も行った．

人名は人名レポジトリ[4]を使用し，語句は n-gram 分割により抽出された語句のうち，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．

役職は，史料中の<役職（氏名＋名前）>というパターンに一致する表現及び史料が作成された時代に対応する官位リストを使用した．

『民経記』では語句の異なり数は 4004，述べ 122023，人物の異なり数は 3360，述べ 54967 だった．

4.2　構造化する機能

史料中の役職表現がどの人物を指すのかを明確にするために文書の構造化を行う．4.1 で選定した役職と人物を結びつけ，人文学資料の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

System development of Topic Estimation for Historical Materials Based on Requirements Analysis of Japanese Historians

†1Katsuya Torii, Department of Systems Innovation, The University of Tokyo

†2Statoru Nakamura and Yamada Taizo, Historiographical Institute, The University of Tokyo

†3Kazuo Hiekata, Graduate School of Frontier Sciences,The University of Tokyo

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.

造化を行った．

4.3 解析する機能

本稿では，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 LDA の解法にオンライン変分ベイズ[5]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 とベクトル表現された史料である corpus が必要であり，4.2 で作成した構造化された史料から dictionary を，dictionary と構造化された史料から corpus を作成した．また，学習中に corpus を通過する回数である passes については，本稿では default 値として perplexity が凡そ収束する 50 を設定した．

4.4 関係性を築く機能

LDA を実行後の出力を変形し，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 2 はトピック数 11 の場合の時系列変化のグラフである．

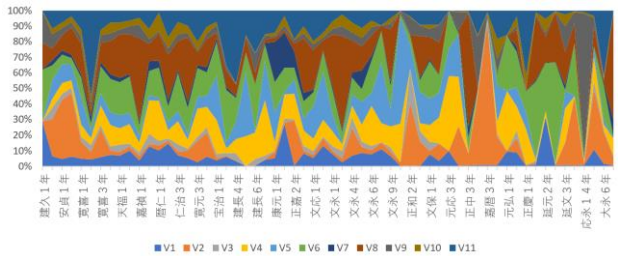

図 2: トピックの時系列変化の積み上げグラフ

4.5 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割合/史料ごとのトピック割合を表示する 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，史料中の単語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，トピック数をユーザーが設定できるように実装した．

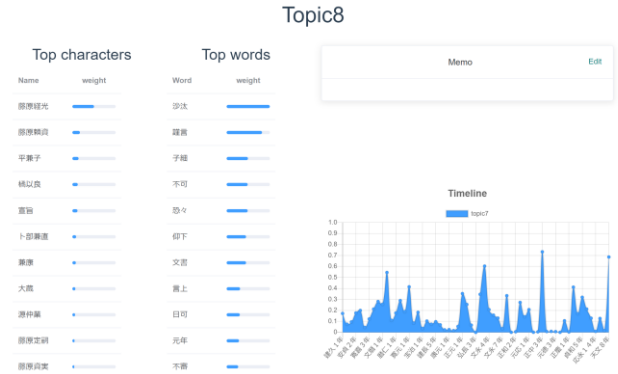

図 3:UI の例.語句と人物のリストとトピックの時系列変化

5．評価

例としてトピック数を 11 とし，LDA の結果の検証を行う．表 1 は 11 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．Topic2 は暦注に関するトピックであり，人物からは筆者の家族及び天皇が関係していることがわかる．Topic3 は神仏関係のトピック，Topic6 は政務関係のトピックである．その他のトピックでも同じように有用な意味を検出することができた．

表 1:トピックと関連する語句及び人物
橙:筆者の親族 ピンク:皇族 黄:政務 青:神仏関係 緑:暦注
灰:地の文 白:判断できないもの

	Topic2		Topic3		Topic6	
	人名	語句	人名	語句	人名	語句
1	幸清	神吉	定玄	禰宜	藤原道家	沙汰
2	藤原頼資	沐浴	相円	供養	藤原泰通	仰下
3	藤原兼頼	天晴	藤原宗氏	導師	藤原家実	奉行
4	愷子内親王	天恩	藤原兼仲	鳥羽	安倍国道	祇候
5	盛家	月徳	藤原宣実	次第	藤原隆親	参内
6	藤原経光	拝官	盛重	其後	藤原有長	相触
7	藤原佶子	歳徳	大中臣隆蔭	御願	藤原兼高	御方
8	藤原資定	母倉	宣旨	下行	藤原忠高	下知
9	仁明天皇	裏書	鳥羽重久	承久	平有親	退出
10	大中臣公行	大小	大中臣隆通	官使	中原俊職	奏聞

6．おわりに

本稿では，人物と語句を Bag-of-Words の対象とし『民経記』の LDA 分析を行った．結果を語句と人物で分け，時系列グラフを作成することでトピックを効率的に理解することを可能とした．また，結果を表示する Web Application の設計を行った．

今後の課題として，『民経記』だけではなく，他の史料にも応用できる柔軟なシステムの開発があげられる．また Voyant Tools や KHcoder などの既存の分析支援システムとの差別化を図りながら，日本史学者の要求を満たすような UI の開発を目指す予定である．

謝辞

本研究の一部は JSPS 科研費 18H03576 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]. 山田太造 , 野村朋弘 , 井上聡 :トピックモデルを用いた天正期古記録『上井覚兼日記』における人物間関係の検出, じんもんこん 2014 論文集, vol.2014, no.3, pp.131-138(2014)

[3]. D. M. Blei, A. Y. Ng, and M. I. Jordan: “Latent Dirichlet Allocation,” Journal of Machine Learning Research, vol.3, pp.993-1022(2003).

[4]. 山田太造 , 遠藤珠紀 , 荒木裕行 , 井上聡 , 久留島典子「前近代日本史史料から人名を集める」, じんもんこん 2016 論文集, Vol.2016, pp.159-164, 2016 .

[5]. 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.

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.