日本南北朝期史料を対象とした潜在的トピックによる史料分類と関連史料提示の手法

山田 太造　　野村 朋弘　　井上 聡
東京大学 史料編纂所　京都造形芸術大学 芸術学部　東京大学 史料編纂所

本論文では，日本南北朝期史料についてテキストを用いた分類および関連史料提示の手法について述べる．史料分類では潜在的トピックを用いる．潜在的トピックを検出するにあたり LDA(Latent Dirichlet Allocation)と呼ばれるトピックモデルを用いる．また，対象史料における潜在的トピックの状態，および史料間の関連について述べる．さらに，潜在的トピックを用いた関連史料提示手法について示す．

A Method of Classifying and Relating Historical Materials in Nanboku-cho Period of Japan

Taizo Yamada
Historiographical Institute
The University of Tokyo

Tomohiro Nomura
Faculty of Art and Design
Kyoto University of Art and Design

Satoshi Inoue
Historiographical Institute
The University of Tokyo

In the paper we introduce a classification method for historical materials in Nanboku-cho period in Japan and indication method of relation among these materials. The classification is realized by latent topics. In order to detect latent topics, we use topic model called LDA (Latent Dirichlet Allocation). We also describe appearances of latent topics in the materials and relation among the materials. Furthermore, we show a method of indicating a historical material related to other with latent topics.

１．はじめに

現在，日本史研究に関わるデータベースの公開が急速に行われており，史料目録・画像・索引だけでなくテキストについても公開が進みつつある．テキストデータベースは，本文に対する文字列一致検索や KWIC（keyword in context;文脈付き索引）などの機能を提供しており，検索クエリのヒットした箇所の前後の文字列を他の史料との比較することができる．しかしながら本文の内容に応じて関連する史料を提示する機能や同じ話題で史料を分類する機能を提供するデータベースはほとんど無い．このような機能があればより本質的に必要な史料を検索結果として得ることができ，さらには史料批判など日本史に関する基礎的研究を促進することができると考えられる．

史料例として『花營三代記』永和元年 9 月 14 日条（群書類従本）をあげる．その記述は次のとおりである．

十四日去八月廿六日午剋於肥後國軍陣太宰少貳冬資爲探題今川伊與入道被誅之由使者到來

先月 8 月 26 日に太宰少貳冬資が今川伊與入道に肥後國の軍陣にて誅されたことを使者から聞いた，という内容である．また，『北肥戰誌』覺書四には次のような記述がある．

冬資討死并今川兄弟所々軍之事（中略）懷良親王ノ令旨ヲ申給テ既ニ宮方ト成リ筑後國へ打テ出味方ノ輩ヲ相催シ（中略）其身ハ肥後國へ打越へ水島ノ城へ楯籠ル因茲探題今川入道六月下旬肥前ヲ出テ肥後國へ發向シ（中略）冬資合戰利ヲ失ヒ籠城叶難カリシカハ八月廿六日生年三十九ニテ竟ニ腹切テ死ス斯テ今川入道ハ水島ヲ攻落シ十二月初筑後へ立歸リ（中略）斯リシカハ了俊ハ頓テ肥前へ歸 陣ス

この史料は，今川入道が水島ノ城に楯籠る冬資を攻落したことが記述されている．先にあげた『花營三代記』に比べ詳細が記述されているが，記述内容としては同様であると考えられる．しかしながら，どれくらい同様であろうか．日本史研究者であれば，時代背景やこれらの史料，および他の史料と比較し，史料批判等を行うことなど，史学的見解により導くことができるかもしれない．その研究結果の１つとして，「北黨今川貞世、肥後水島ノ陣ニ、少貳冬資ヲ誘殺ス、」[1]を導き，この事項のもとでは同様の内容であると結論付けることができよう．一方で，定量的に判断することでこのような結論を導く，もしくは導くための支援を行うことは可能であろうか．また，定量的な判断を行うことができる場合，本当に同様の内容であると結論付けることができるであろうか．

本研究では，日本史史料，特に南北朝期史料に対する，テキストをベースとした史料の分類およ

[1] 『大日本史料』6 編 44 永和元年八月二十六日 2 条綱文

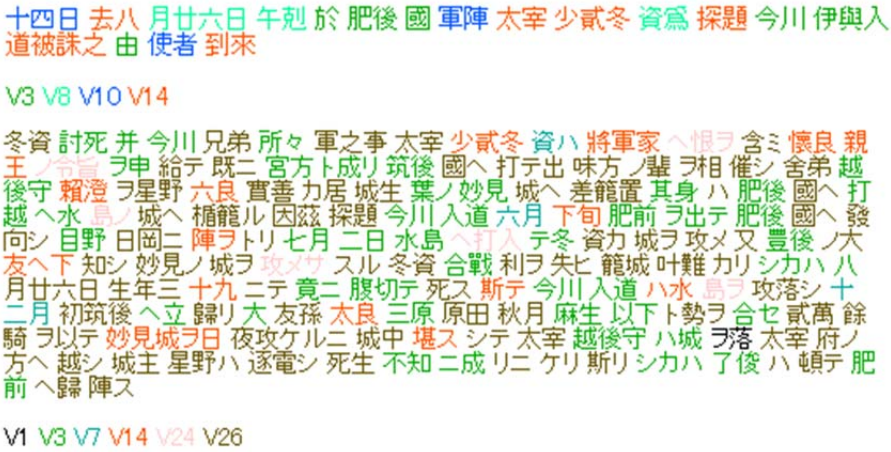

図 1：LDA による用語分類例
Figure1: Example of term classification by LDA
（上）『花營三代記』永和元年 9 月 14 日条（群書類従本）
（下）『北肥戰誌』覺書四

び関連史料を提示する手法を導入する．史料分類ではテキスト内の用語の共起関係をもとに潜在する意味関係を検出し，これにより史料を分類する．ここで，潜在的意味解析するにあたり LDA（Latent Dirichlet Allocation）[1]と呼ばれるトピックモデルを用いる．LDA では 1 史料から複数の潜在的トピック（表層的に現れるのではなく，潜在している話題）が生成され，さらに 1 トピックから複数の用語を生成する文書生成モデルであるため，潜在的トピックの検出，および，潜在的トピックと用語の関係，潜在的トピックと史料の関係を確率的に明示することができる．

例として先の 2 つの史料に対して LDA を適用し，各用語を分類した結果を図 1 に示す．LDA 適用により，『花營三代記』からは 4 つの，『北肥戰誌』から 6 つの潜在的トピックが検出された．そのうち 2 つは共通する潜在的トピック（トピック V3 およびトピック V14）である．2 つの史料とも，トピック V3 およびトピック V14 の潜在的トピックを含む割合は大きいため，共通する内容であると判断できる．他方，『北肥戰誌』は『花營三代記』では検出されない潜在的トピック（主に V26）を含むこともわかる．このように定量的に史料間の関連性を把握することが LDA を利用することで可能となる．

本論文では LDA を用いた史料分類および関連史料提示の手法を述べる．構成は次のとおりである．LDA を用いた潜在的トピックの検出方法を 2 節で述べる．LDA は”bag-of-words”利用を前提としているため，史料テキストを用語分割する必要がある．この方法についても述べる．また，南北朝期史料における潜在的トピック検出に関する実験を行い，その結果を 3 節で示す．さらに，潜在的トピックを用いたテキスト検索システムのプロトタイプについて 4 節で述べる．5 節では本論文での史料分類および関連史料提示の手法について考察し展望を述べる．

２．潜在的トピックの検出

２．１．LDA

文献史料には，記述された内容に何らかの話題がある．多くの場合，これらの話題は史料内に明記されておらず，潜在しており，意味的には読解することで把握することになる．ここではこの潜在する話題を潜在的トピックと呼ぶ．潜在的トピックを検出し，トピックに応じて史料を分類する．潜在的トピック検出のため次式で表現される LDA[1]を用いる．

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

ここでα, βはパラメータ，$z = z_1, z_2, \ldots, z_C$は潜在的トピック，$\theta = \theta_1, \theta_2, \ldots, \theta_C$は潜在的トピックの生成確率，$Dir(\theta|\alpha)$はディリクレ分布，$\mathrm{d} = (\mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_{|d|})$は史料，$w_n$は用語，|d|は史料dの総用語数を示す．LDA は潜在的トピックの生成確率がディリクレ分布に従うと仮定した文書生成モデルといえる．つまり，LDA では 1 史料におけるトピックは複数あり，それぞれのトピックは複数の用語を生成するため，(1) 式を計算することにより各史料の各用語に属するトピックが割り当てられる．各用語をトピックごとに集計することで，各史料のトピック，および全史料におけるトピックの状況を把握することが可能となる．

(1) 式をそのまま計算することはかなり困難であるが，崩壊形ギブスサンプラを用いた解法が知られており[2]，本研究ではこれを用いて潜在

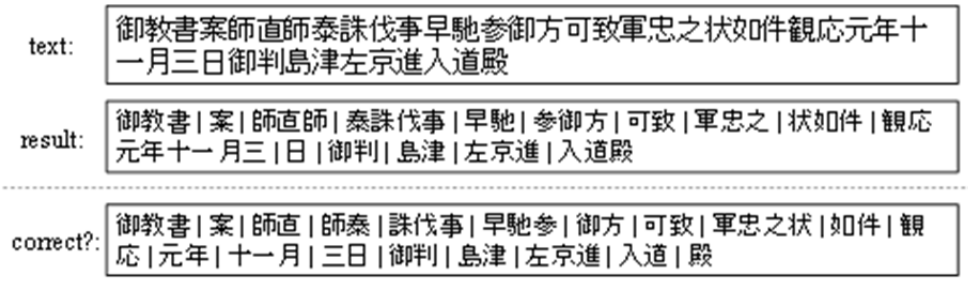

図 2：用語分割例

Figure2: Example of term segmentation

的トピックを算出する. 算出方法の解説については文献[3]を参考にされたい.

トピックモデルとして LDA 以外にも LSI（Latent Semantic Indexing）[4] や pLSI（probabilistic LSI）[5]がある. LSI は 1 史料につき 1 トピックを仮定するため, 多角的な関連性を考慮できない. pLSI は LDA と同様に LSI を拡張し 1 史料につき複数トピックを仮定する. しかしながら, 潜在的トピックの生成確率, つまり（1）式における$p(z|\theta)$を最尤推定するなどして事前に算出する必要がある. そのため, 学習データにはない史料への対応は高コストになってしまう. また$p(z|\theta)$は学習データの量に応じて計算コストが増大してしまうためアドホックな手法で求めることが多い. これに対し LDA は$p(z_k|\theta)$を確率的に算出する生成モデルである.

２．２．用語分割

LDA では潜在的トピックに従って用語が生成され, その結果が史料テキストとして出力されることを示しており, これをもとに学習し, LDA で表現するモデルを推定していく. これを実現するためには, テキストを用語分割する必要がある. 本研究では日本南北朝期史料を対象としている. 一例として「足利直義御教書案」(『島津家文書』)（図 1）をあげる. このような古文書テキストから用語を抽出するのは非常に困難な問題である. 理由としては, 日本語の古文書や古記録などを対象とした形態素解析器がほとんど無いことがあげられる. 現代文とは文法が異なるため, chasen[1]や mecab[2]などの形態素解析器をそのまま用いることは困難である. 形態素解析用辞書の問題もある. 文献[6]のように古典本文に対する形態素解析用辞書の開発が進められているが, 残念ながら, 漢文体, かな文体など, 文体が不均質であるような古文書・古記録への適用はまだ困難な状態にある. 我々は計算機処理に耐える日本に関連する人名や地名に関する辞書を持っていない. 一般的に公開されている各種辞書があるものの, すべての人名や地名などを網羅したものは存在しない. そこで文献[7]の手法を用いて用語分割を行う. この方法は NPYLM と呼ばれるノンパラメトリックベイズ手法にもとづく n-gram 言語モデルを用いて, MCMC 手法と動的計画法により用語らしさを計算し, 推定していく.

適用の具体例は文献[8]を参考にされたい. この手法を用いて用語分割した結果は図 2 に示すとおりである. この結果より, 一部正確とはいえない用語分割（特に“師直師泰誅伐事”を“師直師|泰誅伐事”と人名分割を誤っている箇所）もあるが, 正解と思われる用語分割に近い結果が得られていると考えている. 一方, 図 1 での用語分割では人名や日付等でいくらか正確ではない例が見受けられる. また, 共通して, 日付部分の用語分割は, “日”や元号の直後や“ト勢ヲ”のように助詞が名詞と連結してしまうなど, うまく分割できない場合もある. 特に, 多くの分割失敗は助詞と名詞の連結である. しかしながら, 意味上での分割失敗は多くないため総じて満足できる結果であると思われる. また, 人名や地名も, 辞書を用いていないのにもかかわらず, 分割できていると思われる. さらに, “如件”のような古文書における常套句も問題なく分割できていた（図に示していない）. ただし, 人名“太宰少貳冬資”と“冬資”では分割が異なり, 前者では“太宰”, “少貳冬”, “資”と分割してしまう. これは学習データとして利用した史料テキストにおける文字の統計的特徴によるところが大きいと考えられる.

３．南北朝期史料の潜在的トピック

実際の史料テキストに対し LDA による潜在的トピックの検出を試みた. 日本の南北朝期（元弘 3 年～明徳 3 年（1333～1392））の史料を対象とした. テキストは東京大学史料編纂所データベースにおけるフルテキストデータベース（大日本史

[1] http://chasen-legacy.sourceforge.jp/

[2] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
1	禰宜之	大津	入道々	御房事	コノ一行挿入セラ	一々	一々	謹言上	阿闍梨御房上下二	如件也	住人百	下村事	のあ	親王了	一々
2	神主十	入寺供	合戦之	執金	号文書ト同筆ニカ	不可也	百文之	恐クハ	僧都也	執金	房寺務之	覚円内	へけ	僧正之	二〇
3	足利将軍家執事高	阿闍梨御房上下二	日々	如件也	文書トシテ収メタ	本所之	同々	之由事	法印也	依会	有識免一宇中摂里	行清寄附徳禅寺之	なり合	別当不知山宿禰種	已上一
4	度例	法師三	行揖	仍うりけんの状如	年預之	住持之	三々	之閾	権少僧都法眼和尚	雑掌之	垣内券	観空状	にあ	母に	作人在井十郎百四
5	第二二	年預之	元弘三年七月十七	謹下	五師代	前机	十二月中可致沙汰	一々	権大僧都法眼和尚	在判下	入寺免一宇下居又	次徒弟院事	ともか	永和三年丁巳九月	五斗也
6	義満公	入か	御奉書	候了	アリシ	同前一	二〇	候あ	権憩勲之	可被下	有妻作人	左中将	はあ	座主了	一反三
7	九号文	寺々	進上之	僧正之	ナリト	吉村	三十ケ日	候哉い	護摩師	状如此候	八腸三	田村々	をあい	元年さんようの事	三斗一
8	尊氏判	一自	状如此候	如此之	コノ文書ノ差出者	二〇	五十八	御々	御々	殿いれ候	反定田五斗真長房	花幔	よりて	僭乎	五々
9	参重陸文	集会事	如件也	天氣	第一三	可為三分之百人		可被下	僧正之	之理	四腸持之	名田庄内田村下村	しあ	検校事	三々
10	文歟	行事也	同々	之由事	ホア	新条	五十枚初後夜導師	之処二	金剛定院御	左衛門大尉中原朝	六興	鴇鳴保	しやうくたんのこ	女子に	分米一

	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30
1	一反三	和尚也	状如此候	諸国々	倭将	一々	日本与本省所籍高	一宇寺	御厩	一々	ヨリテ	寺之	敬白奉	次々	時真之
2	百歩ハ	答曰道	寄進セラレタルモ	当時ノモノニアラ	授テ	不可也	本せ	次二十	御薬	之閾	ヲそ	ケねん	陀羅尼田	朝臣也	正吉令
3	反小五	禅師也	右件々	朝臣也	聴聞等	状如此候	曰ク	まき	神田事	一通不可為ヘタ他筆者		一帳	真言之	云々者	正弘一期之
4	料田三	進人	者也了	不可有一紙半銭之	宮主之	年貢之	不膝行	祝言過て御	イせ	於ハ	シテ書	権別当法印大和尚	光明寺	今日之	藤与
5	十歩下田二反六十	云々	四至内	依違之	一職	請文之	建武三	禰宜之	十七	為一	ニあて	和尚位良弘聖俊房	本尊之	晴儀	時友之
6	反半也	不一	限永代所売渡在地	条々	元年	百姓等不引田坪候	沛艾	庁屋二	上分三	之由事	キリ	若宮之	然則下	中納言兼侍従藤原	時沢名文書等之由
7	二反三	如何候	相副之	不論国	六支一	時者也	実無	供を	六也	上者也	ヲハ王	都聞	地蔵堂	大納言家御雑事用	屓物代
8	大々	為一	仍為後日うりけん	遠近之	刷説	可致之	於時真者	社務之	二丁未年	沙汰之	トラス	二内	然際於	御々	彼名者
9	小々	諸仏事	奉寄進之	以後之	又寺	請申也	天皇の	次二日	外宮上	之果	ナカノ	黒田	阿弥陀三	之閾	以御崇敬異于他之
10	畠か	以て	田地一反永代可有	藤原為光重代相伝	次後取々	起請符文	戌時御	持杖	丁五段	可被下	ナリト	三帰	一切可	退出之	状人

図 3：潜在的トピック
Figure3: latent topics

料総合 DB，古記録フルテキスト DB，古文書フルテキスト DB，平安遺文フルテキスト DB，および鎌倉遺文フルテキスト DB）から抽出し，7,007 史料，文字の異なり数が 4,067，延べ文字数が 1,204,594 だった．2.2 節による用語分割により用語の異なり数は 114,576 だった．LDA でのパラメータとして潜在的トピックを 30，ギブスサンプリングの回数を 200 とした．

図 3 は検出した潜在的トピックと各潜在的トピックに関係する用語（出現頻度上位 10 件分であり，この限りではない）を示す．トピック V5 は『大日本史料』における連絡按文（史料の所蔵先や関連する史料などの情報を記述したもの），トピック V8 は古文書，トピック V29 は古記録を示すだろうと思われ，意味的な分類よりも様式や文体に依存していると考えられる．これに対し V1，V3，V9 などはより意味的な分類によるものだと考えられる．例えば，トピック V1 は足利氏に関連する用語が含まれており，御教書関連，もしくは足利氏と深く関係する事項だと推測できる．トピック V3 は今川氏関係，トピック V9 は寺・僧都関係と推測できる．

実際には，それぞれのトピックがどのような特徴を持つかを一言で表すことは必ずしも容易ではなく，個別の史料を確認していく上で把握することができると考えられる．そこで，個別の史料を分析する．図 1（上）は『花營三代記』永和元年 9 月 14 日条，同図（下）は『北肥戰誌』覺書四における各用語の潜在的トピックを示す．LDA によるトピックモデルでは，ある 1 用語は複数のトピックに属することも可能である．用語がどのトピックから生成されるかは史料ごとに異なる．例えば，この 2 つの史料とも“太宰”が出現する．しかしながら，『花營三代記』ではトピック V14，『北肥戰誌』ではトピック V26 に属している．これは“太宰”と共起する用語との関係によるものと推測できる．つまり，2 つの史料とも，“太宰”と“少貳冬資”が共起するが，『北肥戰誌』では“太宰”と“府”が共起している．前者は人名，後者は地名を示す．意味的に異なるため，トピックが異なる結果となったと考えられる．

図 4 は『花營三代記』永和元年 9 月 14 日条における各用語に対して，この実験で用いた全史料における用語がどのトピックから生成されたかを示す．“今川”の全史料中での出現頻

	今川	伊與入	使者	到來	十四日	午剋	去八	國	太宰	少貳冬	探題	於	月廿六日	由	肥後	資爲	軍陣	道被誅之
V1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V2	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
V3	33	3	1	0	0	0	1	12	0	1	13	182	10	38	29	0	0	0
V4	0	0	0	0	23	0	0	14	0	0	0	0	10	152	0	0	0	0
V5	0	0	11	0	0	0	0	0	0	0	0	0	7	0	0	0	0	0
V6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V7	0	0	0	0	26	0	0	0	0	0	0	0	27	0	0	0	0	0
V8	0	0	17	8	25	2	0	0	0	0	5	53	28	44	0	1	0	0
V9	1	0	0	0	21	0	0	0	0	0	0	74	0	0	0	0	0	0
V10	0	1	6	0	34	0	0	0	0	0	0	67	3	13	0	0	4	0
V11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V13	0	0	0	0	7	0	0	0	0	0	0	0	0	0	1	0	0	0
V14	0	0	0	3	0	0	1	6	8	3	13	8	1	1	0	0	0	1
V15	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0
V16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V17	0	0	3	0	0	0	0	0	0	0	0	30	0	0	0	0	0	0
V18	0	0	0	0	9	0	0	0	0	0	0	103	18	0	0	0	0	0
V19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V21	0	0	0	0	5	0	0	0	0	0	0	44	0	0	0	0	0	0
V22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V23	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	0
V24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V25	0	0	27	0	3	0	0	0	0	0	0	659	17	153	0	0	0	0
V26	0	0	0	0	0	0	0	0	6	0	14	0	2	0	0	0	0	0
V27	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
V29	0	0	0	0	22	1	0	0	0	0	0	101	3	11	0	0	0	0
V30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

図 4：潜在的トピックと用語の関係

Figure4: relation between latent topics and terms

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
花營三代記	0	0	0.33333	0	0	0	0.16667	0	0.16667	0	0	0	0.33333	0	
北肥戰誌	0.00671	0	0.30872	0	0	0	0.02013	0	0	0	0	0	0.10067	0	

	V16	V17	V18	V19	V20	V21	V22	V23	V24	V25	V26	V27	V28	V29	V30
花營三代記	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
北肥戰誌	0	0	0	0	0	0	0	0	0.04027	0	0.52349	0	0	0	0

図 5：潜在的トピックの含有率

Figure5: content rate of latent topic

度は 34 であり，そのうちトピック V3 から生成されたのは 33, トピック V9 から生成されたのは 1 であることがわかる．この場合，ほぼトピック V3 から生成されている．そのため，このトピックを特徴づける用語としてふさわしいと考えられる．他方，“十四日”，“於”，“由”などは複数のトピックから生成されている．そのため，各トピックを特徴付ける用語とはいえないかもしれない．

次に『花營三代記』および『北肥戰誌』における各トピックの含有率を図 5 に示す．この図から 2 つの史料ではトピック V3 の含有率は約 30%，トピック V14 の含有率は『花營三代記』では約 30%，『北肥戰誌』では約 10%であり，この 2 つの共通するトピックにより 2 つの史料を関連付けることができる．また，『北肥戰誌』はトピック V26 を約 52%も含有する．この点において 2 つの史料は必ずしも同等ではないことがわかる．図 5 で示すような史料と潜在的トピックの関係から史料の特徴を見出すことが可能であると考えられるため，これを潜在的トピックによる史料の特徴ベクトルとして扱うことが可能になる．これにより，史料テキストを定量的に分析することが可能になり，例えば，史料間の類似度のような関係をこれにより算出することが可能となる．

4．史料テキスト検索システム

本研究において，史料テキストを検索するシステムをプロトタイピングした．本システムは，フルテキスト検索，検索結果の関連史料提示，および関連史料のタイムライン表示の機能を提供する．本研究における関連史料提示手法は LDA を

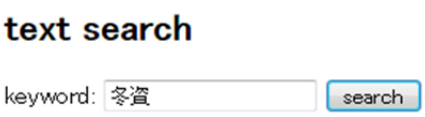

図 6：検索ページ
Figure6: search page

図 7：検索結果一覧
Figure7: search result list

用いて算出した潜在的トピックを用いる．本節は，本手法およびプロトタイピングシステムの概要について述べる．

４．１．史料間の類似度

ある史料に対する関連史料を提示するため，史料間の類似度を用いる．史料間の類似度は潜在的トピックを用いて算出する．(1) 式の計算により各史料の用語の属するトピックが決定される．各史料dにおいてトピックz_kから生成された全用語$w_{k,1}, \ldots, w_{k,|d|}$の出現頻度を総計したものをdにおける$z_k$の出現頻度とする．

次に史料の特徴ベクトルを生成する．史料xにおけるベクトルの各要素をトピックの重み$\mathrm{weight}(z_k, x)$とする．トピックの重みは出現頻度を用い，次式に示す tf-idf 重み付けと同様な方法で算出する．

$$\mathrm{weight}(z_k, x) = \mathrm{tf}(x_k) \cdot \left(\log \frac{N}{df(k)+1}\right) \qquad (2)$$

ここで$\mathrm{tf}(x_k)$はxにおけるz_kの出現頻度を，df(k)はz_kを含む史料の点数，Nは史料数を示す．(2) 式により，各トピックの重みに対して，図 4 における“十四日”や“於”のように多くのトピックから生成される用語の影響を弱め，“今川”のように特定のトピックのみから生成される用語の影響を高めることができる．

史料xと史料yの類似度は次式のコサイン類似度を用いて計算する．

$$\mathrm{sim}(x, y) = \frac{\sum_k weight(x_k) \cdot weight(y_i)}{\sqrt{\sum_k weight(z_{ki}, x)^2} \cdot \sqrt{\sum_k weight(z_{ki}, y)^2}} \qquad (3)$$

４．２．史料テキスト検索システム

本システムは，フルテキスト検索を行い，検索結果をリスト形式表示およびタイムライン表示し，検索結果に応じて関連する史料を提示する機能を提供する．

図 6 は検索ページである．現在の本システムはテキスト内の文字列の検索しか行わないため，かなりシンプルなページである．検索クエリを入力し，search ボタンをクリックすることで検索結果一覧画面へ遷移する．

図 7 は検索結果一覧画面である．図 6 において入力した検索クエリ，検索結果件数，および検索結果を一覧して表示している．このとき，検索クエリを史料とみなし，検索クエリから用語を抽出し，3 節で求めた全史料におけるトピックからの用語生成頻度を求め，それと (2) 式によりクエリの特徴ベクトルを作成する．このクエリ特徴ベクトルと史料特徴ベクトルの間の類似度を (3) 式で求める．この類似度をもとに検索結果をソートしている．

図 8 は各史料の検索結果詳細画面を示す．史料の名称およびテキストとともに，図 1 で示した潜在的トピックごとに用語を色分けして提示する．また，関連史料を提示する．この関連史料は (3) 式により求め，史料間の類似度に応じてソートする．

図 8 から“to timeline”をクリックすることで図 9 に示すタイムライン上に選択した史料とともに関連史料を配置することができる．このタイムラインシステムは Simile Timeline[1]を用いて構築している．ここでは選択したタイムライン上に配置した史料のアイコンの色を類似度に従ったランクに応じて変更している．また，各史料をクリックするとその史料のタイトルやテキストなどを表示する．さらに，表示したタイトルをクリックすると，その選択した史料の関連史料をタイムライン上に表示することができる．

５．考察と展望

本節では LDA を用いた史料分類および用語分割について考察しその展望を示す．

[1] http://www.simile-widgets.org/timeline/

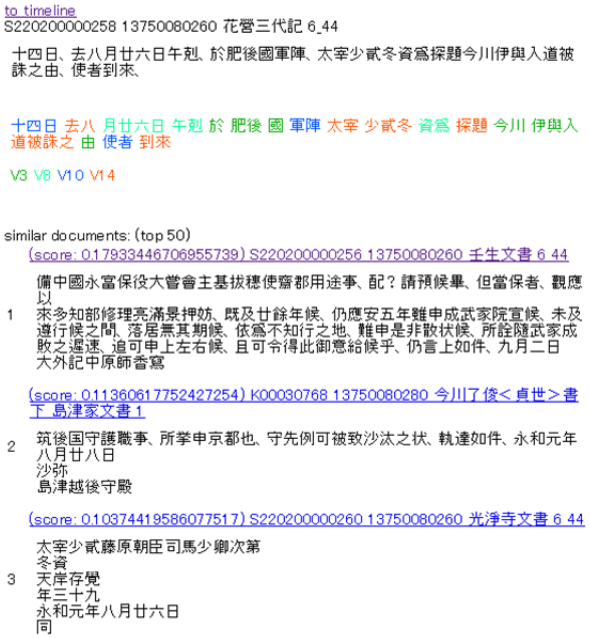

図 8：関連史料提示
Figure8: Indicating related materials

図 9：タイムライン表示
Figure9: timeline

5．1．史料分類

3節で検出した潜在的トピックは一体何であろうか．一般的に LDA などのトピックモデルは，様々なデータやテキストなどから，話題・論題・事柄・出来事のような，明記されていない潜在するトピックの推定を行う．図 1 で示したとおり，『花營三代記』永和元年 9 月 14 日条は 4 つのトピック，『北肥戰誌』覺書四は 6 つのトピックを持つことがわかり，2 つの史料の関係を定量的に明確化することが可能となった．

多変量解析手法としては k-means などのクラスタリング，主成分分析，因子分析などがあり，従来データマイニング・テキストマイニングにおいて利用されてきた．例えば主成分分析では，上位 N 個の主成分のみに着目する．そのため，解析漏れがあるかもしれない．また，各主成分がどのようなどの成分に由来するのかが明確化されない．これに対し，LDA では，ある潜在的トピックに属する用語が史料ごとにわかるため，各潜在的トピックの由来が明確である．また，すべての潜在的トピックについてその状態を把握することができるため，主成分分析のように主成分を限定することはない．これにより，低頻度の用語，潜在的トピックについても分析を行うことが可能である．また，古典的なクラスタ分析，たとえば k-means では，クラスタリングを行う前にクラスタ数を設定する必要がある．また，排他的にクラスタリングするため，あるクラスタに属する史料であってもシード（クラスタの中心）から遠い場合もある．さらに 1 つのクラスタにしか属することができない．これに対し，LDA での分析では，非排他的であり，どのトピックにどれくらい属しているか（各トピックにどれくらい近いか）がわかる．

図 3 よりトピック V8 は古文書，トピック V29 は古記録の書式を示すものと推測できる．つまり，LDA による潜在的トピックの検出を用いることで，書式に依存する用語を分類することが可能であると考えられる．この結果は大まかな分類である．扱う史料をある家の古文書に限定した場合を考える．その場合，公家様文書・武家様文書上申文書・証文のような分類が可能であると考えている．さらに，武家様文書でも下知状・御教書・奉書・直状のようなさらに下位の様式の分類を行うことができるかもしれない．すなわち，古文書様式について定量的分析を行うことができるようになる．

また，家分け文書のみを扱う場合を考えてみる．『島津家文書』は薩摩の島津氏が平安時代末期から明治時代初期までの約 700 年間代々伝えてきた文書群であり，南北朝期においても当時の研究を行う上で重要な史料は多く収載されている．これに対し，LDA を適用することで，古文書様式の時系列変化を分析することが可能であろうと思われる．また，他の同時代の家分け文書との比較を行うことで，日本における古文書様式変化の把握が可能になる．4 節で述べたプロトタイピングシステムでは時系列に史料を配置する機能を提供している．現時点ではある史料に関連する史料のみを配置するのみである．これを発展させ，

ある家分け文書・古記録のみを配置したとする．そうすれば，家分け文書・古記録における潜在的トピック，つまり内容の時系列的変化を可視化できると考えている．

本研究では潜在的トピックを 30 とした．これはアドホックな方法である．対象史料にもとづいて潜在的トピックの数を決めるべきである．適切なトピック数を特定するため，トピック数を変動させ，結果を分析することで決定することも考えられる．他方，トピック数自体を（1）式のパラメータとして算出することも考えられる．

５．２．用語分割

本研究では2.2節で示したNPYLMによる教師なし学習に基づく用語分割手法を用いた．その結果の一部は図 1 および図 2 に示した．特定の人名地名において適切ではない分割も見られたが約50%の確率で正解と思われる分割を行うことができた．2.2 節でも述べたとおり，分割失敗の最たる例は助詞と名詞の連結である．意味的分析を行う場合，名詞の分割失敗に比べ影響は低減すると考えられる．

一方，用語分割の精度を向上することで LDA による史料分類の精度を向上させることができると考えられる．本研究では現時点では人名・地名などの辞書を利用していない．そこで，このような辞書を用いることで，用語分割の精度を向上させることができると考えている．しかしながら，南北朝期に限定したとしても，人名・地名などを網羅した辞書データは存在しない．そこで，限定的ではあるが，整備されている人名・地名の辞書を用い，NPYLM による用語分割手法を拡張し，半教師あり学習による手法の導入を検討したいと考えている．

６．おわりに

本論文では，日本南北朝期の史料テキストに対して，NPYLM による用語分割，および LDA を用いた方法により史料分類を行う手法を示し，史料間の関連性について述べた．さらにプロトタイピングした史料テキスト検索システムの概要，テキスト検索および関連史料提示の機能について示した．

LDA による史料分類手法により，大まかなではあるが取り扱う史料の全体像を把握することが可能となる．さらに，個別の史料を見ていく上で，関連性を持つ史料を把握することができる．このとき，各史料の間で，どの要因により関連するのかを明示しながら確認することができるのが LDA による史料分類の特徴である．今後は更に掘り下げ，史学研究の展開と LDA による史料分類，つまりテキストマイニングの方法を対照することで，その性能をより向上させることを狙う．また，対象史料を戦国期へと拡張させる．戦国期の史料は南北朝期に比べ，その様式も多様であり，史料点数も格段に多くなる．そのため，テキスト分析の利用を最も求めている時代だと考えている．

謝辞

本研究の成果の一部は，日本学術振興会科学研究費基盤研究（A）「ボーンデジタル画像管理システムの確立に基づく歴史史料情報の高度化と構造転換の研究」（23240031）の助成を受けたものによる．

参考文献

1) D. M. Blei, A. Y. Ng, and M. I. Jordan: “Latent Dirichlet Allocation,” Journal of Machine Learning Research, vol.3, pp.993-1022, 2003.
2) T. L. Griffiths and M. Steyvers: “Finding scientific topics,” Proc. of the National Academy of Sciences of the United States of America, vol.101, pp.5228-5235, 2004.
3) 手塚太郎：” LDA (Latent Dirichlet Allocation)の更新式の導出”，入手先〈http://yattemiyou.net/docs/lda_gibbs.pdf〉（参照 2013-08-26）.
4) S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman: Indexing by Latent Semantic Analysis, Journal of the American Society of Information Science, Vol. 41, No. 6, pp. 391-407 (1990).
5) T. Hofmann: Probabilistic Latent Semantic Indexing, Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 50–57 (1999).
6) 小木曽智信，小椋秀樹，田中牧郎，近藤明日子，伝康晴：中古和文を対象とした形態素解析辞書の開発，情報処理学会研究報告,人文科学とコンピュータ研究会報告, Vol.2010, No.4, pp.1-8, 2010.
7) 持橋大地，山田武士，上田修功：ベイズ階層言語モデルによる教師なし形態素解析(言語モデル・ウェブ解析)，情報処理学会研究報告．自然言語処理研究会報告, Vol. 2009, No. 36, p. 49 (2009).
8) 山田太造,「関連史料収集のための手法に関する考察-日本の南北朝期における史料を対象に-」,『研究報告人文科学とコンピュータ（CH）』, Vol.2013-CH-97,no.6,2013 年,pp.1-6.