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

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，天正期古記録『上井覚兼日記』のテキストを用いて，そこに出現する人名に対する関連検出の手法および提案手法を用いてプロトタイピングしたテキスト検索システムについて述べる．人名間関連検出は，トピックモデル LDA（Latent Dirichlet Allocation）により検出された潜在トピックを利用することで実現した．『上井覚兼日記』における潜在トピックの検出に関する実験を行った．その検出結果と時系列変化について述べる．

Detection of Relationship between People Using Topic Model from “Uwaikakken-nikki” as Tensho Period Diary

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 detection method of relationships between personal names using text corresponded with “Uwaikakken nikki” which is a diary of Tensho period and prototyped text search system which utilizes the detection method. The detection method is realized by using latent topics which is detected by LDA (Latent Dirichlet Allocation) as topic model. We experimented with the detection of latent topics in “Uwaikakken nikki”. We explain the experimental results as the detection results and time series variation of the latent topics.

１．はじめに

日本史学研究の基礎は，史料（古文書・古記録・典籍など）を精確に読解し，内容を批判的に検討し，そこから論点を引き出して歴史像を構成する[1]ことである．精確な読解や，批判的検討を行う史料批判等を行う上で，歴史的背景・知識が必要であり，その１つが人名に関する情報ある．

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．史料内に出現する人物や人物の関係を提示していくことは歴史像を見出す上でその解決に非常に有益であると考えられる．史料目録に関するデータベース（国立歴史民俗博物館館蔵古文書データベース[2]，東京大学史料編纂所データベース（以下，SHIPSDB）大日本古文書ユニオンカタログ[3]など）の一部では，史料内に出現する，もしくは史料に関係する人名を提示する仕組みを有する．しかしながら，テキストデータを利用した人物提示や人物間の関係を提示するシステムは我々の知る限りにおいてない．

本研究では，天正期における薩摩島津家の家臣である上井覚兼の日記『上井覚兼日記』を対象に，史料とそのテキスト内に現れる人名を関連付け，史料テキスト内での人名の共起に従って人物間の関係を検出していく手法を提案する．テキスト内の人名の共起関係をもとに潜在する意味関係を検出し，これにより人名間の関係性を見出す．ここで，潜在的意味解析するにあたり LDA（Latent Dirichlet Allocation）[4]と呼ばれるトピックモデルを用いる．本研究においては，LDA により 1 史料から複数の潜在トピックが生成され，さらに１トピックから複数の人名を生成する文書生成モデルとみなすことができる．これにより，潜在トピックの検出，および，潜在トピックと人名の関係，潜在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．LDA は”bag-of-words”の利用を前提としているため，史料テキストから予め人名を抽出する必要があるため，この手法を示す．人名の共起に従った潜在トピック検出に関する実験を行い，その結果を示す．さらに，潜在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も述べる．

(c) Information Processing Society of Japan

DB内での表現	抽出結果	ルール
xx(yy)	xxyy，もしくは yyxx	他の注記，もしくは抽出結果に出現した方を採用
xxz州(yy)	xxyy	“z州”を削除
xx<官職名>(yy)	xxyy	官職名(山城守など)を削除
(義久|義弘|歳久|家久)	島津xx	島津4兄弟の場合は島津を補う
拙者	上井覚兼	日記の著者名で置換

図１：人名抽出ルールの例
Figure1: example of extraction rules of personal name

２．『上井覚兼日記』

『上井覚兼日記』は薩摩島津家の家臣であった上井覚兼による日記である．『伊勢守日記』とも呼ばれる．原本は東京大学史料編纂所（以下，史料編纂所）所蔵の重要文化財であり，国宝『島津家文書』に含まれる史料である．また史料編纂所はこの影写本も所蔵している．原本は 28 冊で構成されている．天正 2 年 8 月 1 日から天正 14 年 10 月 15 日までの約 12 年間の日記であるが，このうち，天正 4 年 9 月 7 日から天正 10 年 11 月 3 日までの日条は存在しない．

天正 2 年 8 月 1 日から天正 4 年 9 月 6 日までは鹿児島における奏者時代の，天正 10 年 11 月 4 日から天正 14 年 10 月 14 日までは日向国宮崎城主時代の日記として見ることができる．沖田畷の戦いや岩屋城の戦いに代表される島津家と龍造寺家・大友家との関係や，羽柴秀吉による九州征伐寸前までの島津家の動向など，天正期における島津家や九州の情勢を知る上で重要な史料として位置づけられている．

『上井覚兼日記』は『大日本古記録 上井覚兼日記』（全 3 冊）として編纂史料集が発刊されている．これをベースとして，SHIPSDB「古記録フルテキストデータベース」（以下，古記録 FT）に，この全テキストが格納されており，テキスト検索サービスを提供している．

３．人名間関係の検出

本節では，本研究における人名間の関係を検出する方法について述べる．

３．１．潜在トピックの検出

文献史料には，記述された内容に何らかの話題がある．こ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トピックと呼ぶ．潜在トピックを検出し，このトピックに応じて史料を分類する．潜在トピック検出のため次式で表現される LDA[4]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，$Dir(\theta|\alpha)$はディリクレ分布，$\mathrm{d} = (\mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_{|d|})$は史料，$w_n$は単語，|d|は史料 d の総単語数を示す．LDA は潜在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．

本研究に当てはめた場合，LDA は，1 史料におけるトピックは複数あり，トピックはそれごとに複数の人名を生成することをモデル化している．(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[5]，本研究ではこれを用いて潜在トピックを算出する．

トピックモデルとして LDA 以外にも LSI（Latent Semantic Indexing）[6]や pLSI（probabilistic LSI）[7]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，潜在トピックの生成確率，つまり式（1）における$p(\mathrm{z}|\theta)$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$p(\mathrm{z}|\theta)$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．

３．２．人名の抽出

LDA を利用する上で，史料からの人名抽出は必須である．本研究では，古記録 FT にて公開されているテキストを利用することにした．このデータベースでは和暦データとテキストを有して

(c) Information Processing Society of Japan

	人名	頻度		人名	頻度		人名	頻度
1	上井覚兼	1377	11	肝付兼寛	40	21	島津歳久	27
2	島津義弘	82	12	鎌田兼政	40	22	本田薫親	26
3	伊集院忠棟	74	13	上原尚近	39	23	本田正親	24
4	島津家久	71	14	伊地知重秀	37	24	上井兼成	23
5	島津義久	70	15	上井薫兼	35	25	吉利久金	22
6	島津忠長	59	16	伊集院久宣	35	26	有馬晴信	22
7	柏原有閑	50	17	伊集院久治	34	27	山有信	20
8	本田親貞	45	18	平田光宗	31	28	新納久時	20
9	比志島義基	43	19	平田宗張	30	29	白浜重治	20
10	吉利忠澄	42	20	喜入季久	27	30	平田増宗	19

図２：人名の出現頻度
Figure2: appearance frequencies of personal names

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	上井覚兼	上井覚兼	上井覚兼	上井覚兼	上井覚兼	上井覚兼	上井覚兼	上井覚兼	上井覚兼	上井覚兼
2	本田親貞	島津義久	吉利忠澄	柏原有閑	島津歳久	吉利忠澄	比志島義基	島津義久	島津義弘	島津家久
3	平田増宗	伊集院忠棟	秋月種実	鎌田兼政	本田薫親	上井薫兼	伊集院久宣	伊集院久治	上原尚近	島津義弘
4	白浜重治	伊地知重秀	龍造寺政家	上井兼成	阿多忠辰	肝付兼寛	吉利久金	島津忠長	島津忠長	伊集院忠棟
5	伊地知重秀	北郷時久	高城珠長	野村豊綱	喜入季久	鎌田兼政	平田宗張	川上久隅	島津家久	三田井親武
6	八木昌信	本田親貞	大友義統	上井薫兼	土持久綱	伊集院忠棟	新納久時	島津義弘	島津義久	上井神九郎
7	肝付兼寛	喜入季久	島津歳久	桑幡道隆	敷祢頼賀	有馬晴信	大寺大炊助	伊集院忠棟	山忠助	上井秀秋
8	伊集院忠棟	島津以久	稲長辰	田尻鑑種	島津以久	大寺大炊助	川上翌久	山有信	名和顕孝	柏原有閑
9	島津忠長	平田昌宗	有馬晴信	平田宗張	川上久辰	島津義久	山忠助	本田正親	平田歳宗	有川貞真
10	平田光宗	島津義弘	筑紫広門	敷禰頼元	平田光宗	平田光宗	吉利忠澄	本田親正	新納久時	本田正親
11	本田薫親	島津朝久	上原尚近	土持久綱	赤星統家	島津忠長	山有信	合志親重	伊集院久治	税篤和
12	吉田清存	伊地知重隆	井尻祐貞	平田宗応	宮原景種	平田宗衡	上井秀秋	村田経平	伊集院忠棟	島津忠長
13	長谷場純辰	平田宗張	島津義弘	新納忠真	島津家久	敷祢頼賀	吉田清存	猿渡信光	大野忠宗	志岐鎮経
14	税篤和	村田経定	平田宗位	三池鎮実	有馬晴信	鎌田政景	吉利忠富	喜入季久	新納教久	新納忠包
15	龍造寺隆信	東郷重尚	伊地知重則	三田井親武	稲長辰	奈良原延	大寺政安	島津忠良	三原重益	本田親商
16	上原尚近	河野通貞	鎌田政虎	市来忠末	本田盛親	島津家久	小川有季	新納久饒	北郷忠虎	野村友綱
17	木脇祐充	国分定友	織田信長	桂忠詮	本田親兼	肝付兼篤	市来家守	川田義朗	宮内忠増	島津久保
18	羽柴秀吉	肝付兼寛	近衛前久	平田宗清	猿渡信光	伊地知重元	柏原有閑	小代親泰	山有信	伊集久治
19	三原重隆	伊地知重興	伊集院忠棟	猿渡信光	鎌田政景	大山綱秀	比志島国貞	川上忠智	島津歳久	北郷久盛
20	島津貴久	本田親治	志岐親重	猿渡信孝	高城珠長	川上忠智	稲長辰	平田光宗	本田正親	吉利久金

図３：潜在トピックと関連する人名
Figure3: personal names related with latent topics

いる．テキストは XML 形式等で構造化されていないが，人名に対して注記が付与されている．例えば「天正 14 年 6 月 7 日条」は次のとおりである．

（前略）…候、従武庫様（島津義弘）急ニ御申之義候間、昨日より御談合にて候、伊地知伯州（重秀）・吉田作州（清存）にて様子承候、…（後略）

このようにカッコ書きで注記が表現されているため，該当箇所を抽出していくことで人名と史料を関係付けていく．この例では，“島津義弘”は容易に抽出することができる．しかしながら，同様の方法で他の人名を抽出すると，“重秀” や “清存”となり，家名を抽出するに至らない．そのため同名の人物が出現した場合への対応が不十分である．他にも同日の日条には以下の記述がある．

（前略）…忠長（島津）・川上上州（久隅）・忠棟（伊集院）…（後略）

そこで本研究では，古記録 FT における人名出現におけるパターンを見出し，それに応じて抽出のためのルールを設定し，自動的に抽出することにした．抽出ルールの一部を図 1 に示す．

３．３．人名 bag-of-words の作成

テキストの単位ごとに，2.2 節での人名抽出方法を適用することで人名を抽出し，そのテキスト内での出現頻度を用いて bag-of-words を作成する．例えば，「天正 2 年 8 月 4 日」条であれば，“喜入”，“東郷”，“拙者”，“白浜周防介”が人名

(c) Information Processing Society of Japan

	島津義久		島津義弘		島津歳久		島津家久	
1	0.93946	本田親正	0.99499	本田親豊	0.99756	本田親兼	0.91130	山善久
2	0.92076	川上久隅	0.87742	深水宗方	0.98432	川上忠智	0.88378	柏原有閑
3	0.89796	島津忠良	0.84678	島津彰久	0.93050	野村重綱	0.79634	田尻鑑種
4	0.89059	大津山家稜	0.80232	島津忠長	0.88588	伊地知重次	0.79343	上井神九郎
5	0.89059	渋谷松右衛門尉	0.78772	高城珠長	0.87977	税篤和	0.79343	野村豊綱
6	0.89059	辺春親行	0.77243	大津山家稜	0.86777	本田薫親	0.79261	上井兼成
7	0.88656	本田正親	0.77243	渋谷松右衛門尉	0.84289	本田親正	0.75798	比志島義興
8	0.88400	伊集院久春	0.77243	辺春親行	0.80524	大津山家稜	0.74972	平田宗応
9	0.88164	村田経平	0.76466	村田経平	0.80524	渋谷松右衛門尉	0.74552	鎌田兼政
10	0.88133	山有信	0.76067	山有信	0.80524	辺春親行	0.73595	近衛信尹
11	0.87464	町田忠綱	0.75076	島津義久	0.80485	平田歳宗	0.66292	名和顕孝
12	0.86714	伊集院久治	0.74521	猿渡信光	0.79715	村田経平	0.63726	吉利忠澄
13	0.86572	猿渡信光	0.74236	川上久隅	0.79299	山有信	0.59251	上井薫兼
14	0.82731	宮原景種	0.73279	伊集院久春	0.78090	伊集院久春	0.58867	伊地知重元
15	0.82689	比志島国貞	0.73137	高城珠玄	0.78086	平田光宗	0.55893	有川貞真

図４：島津 4 兄弟との共起関係
Figure4: co-occurrence features for Shimadu brothers

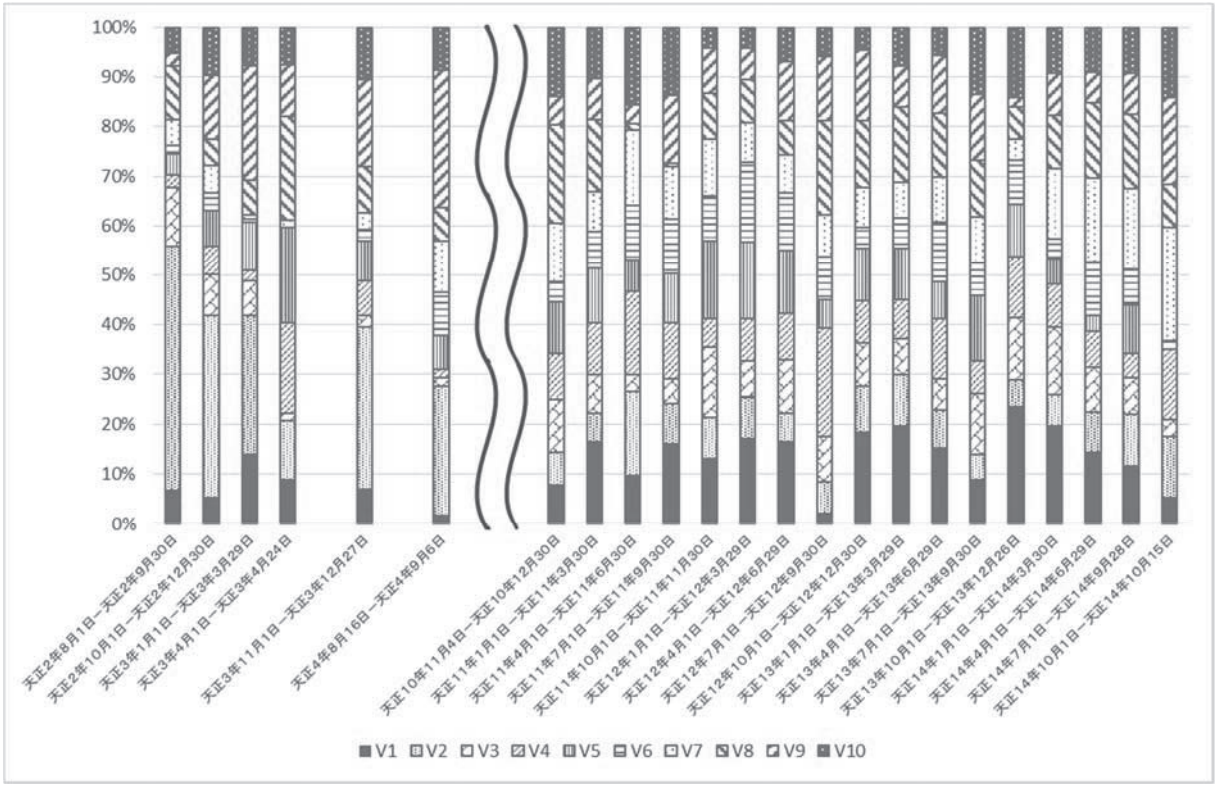

図５：潜在トピックの時系列変化
Figure5: time series variations of latent topics

として認識され，人名抽出のルールに従って，それぞれ，“喜入季久”，“東郷重尚”，“上井覚兼”，“白浜重政”が抽出される．またそれぞれの人名の出現頻度は 1 回であった. これを bag-of-words で表現する．

このとき，テキストの単位は古記録 FT と同様に段落とした．出現する人名は，本テキストデータ内では同姓同名は確認されないと思われるため，人名における文字列の異なりを人名の識別に用いた．

4．実験

『上井覚兼日記』に対し LDA による潜在トピックの検出を試みた. 対象となる史料テキストは，目録（日条，および段落）単位で 1777，人名の異なり数は 520 だった．LDA でのパラメータとして潜在トピックを 10，ギブスサンプリングの

(c) Information Processing Society of Japan

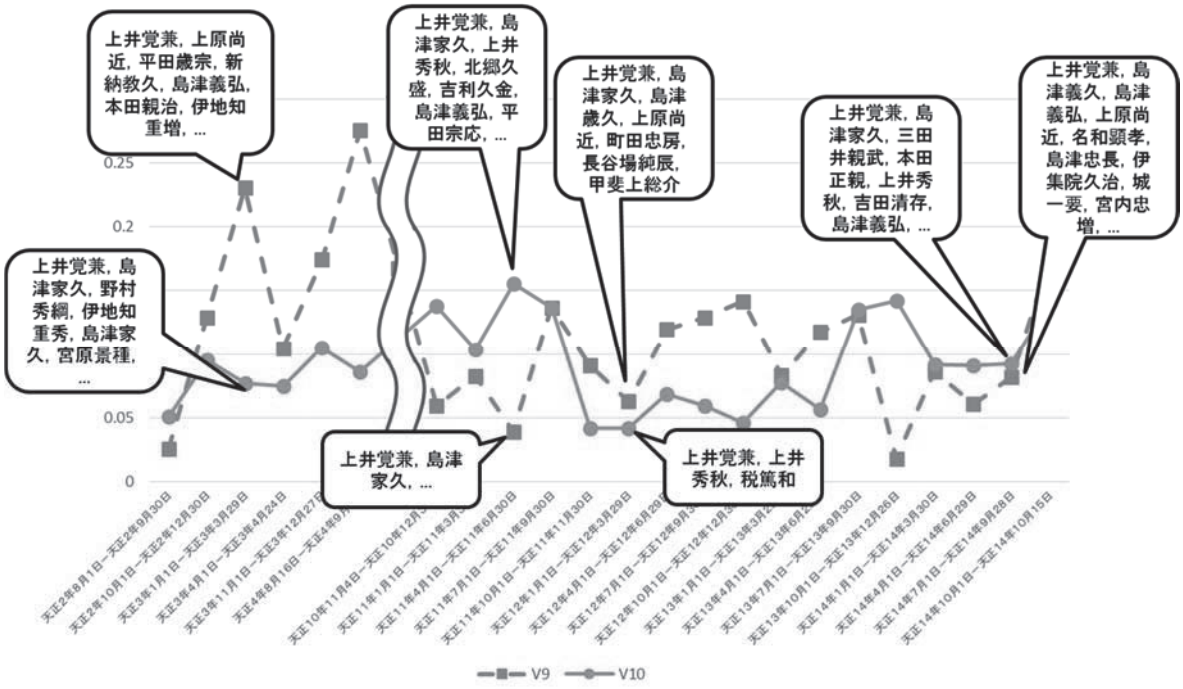

図６：V9 と V10 における潜在トピックの時系列変化
Figure6: time series variations of latent topics for V9 and V10

上井覚兼日記検索システム

q="家久" --> 件数:207

ID	和暦	段落番号	本文	画像
00035171	天正2年8月16日	16	一、十六日、如常出仕候、入来院殿血判上候、同内衆五人之血判も	画像
00035173	天正2年8月18日	18	一、十八日、出仕如常申候、中書様御内儀老中衆迄被仰候、此儀拙	画像
00035210	天正2年9月26日	55	一、廿六日、如常出仕申候、野村美作守殿より之書状之趣、御老中	画像
00035213	天正2年9月29日	58	一、廿九日、如常出仕申候、伊地知式部太輔殿より、同名源左衛門	画像
00035263	天正2年11月19日	108	一、十九日、如常出仕申候、和光院へ一昨日之返事申候、一、此	画像
00035264	天正2年11月20日	109	一、廿日、出仕前に中書様へ、坂本二郎兵衛面目失せられ候由申候	画像
00035283	天正2年閏11月9日	128	一、九日、如常出仕申候、中書様より御老中迄と候て御申候、隈城	画像
00035303	天正2年12月1日	148	拾二月 一、朔日、如常出仕申候、従和泉伊地知善介と申使者被参	画像
00035312	天正2年12月10日	157	一、十日、如常出仕申候、従泰平寺御申候、彼むかへに立花團と申	画像
00035314	天正2年12月12日	159	一、十二日、如常出仕申候、中書様より御申候、此前委被聞召候、	画像
00035316	天正2年12月14日	161	一、十四日、中書様於御仮屋御連歌候、其人数に参し候間、常之出	画像
00035318	天正2年12月16日	163	一、十六日、如常出仕申候、河内誾寺家歳暮之御祝言に参候、取成	画像
00035319	天正2年12月17日	164	一、十七日、如常出仕申候、河上犬万殿懸御目候、式三献にて候、	画像
00035338	天正3年1月6日	183	一、六日、片衣袴にて出仕申候、不断光院・浄光明寺御旨出候、御	画像
00035365	天正3年2月3日	210	一、三日、如常出仕申候、串木野へ御使被仰付候、明後日五日彼方	画像
00035367	天正3年2月5日	212	一、五日、串木野へ参候、酉刻末に中書御前二罷出候、驚而御意趣	画像
00035368	天正3年2月6日	213	一、六日、小宿へ中書様御出にて、御返事被成候也、此日、冠嶽へ	画像

図７：検索結果一覧
Figure7: search result list

回数を 2000 とした.

４．１．トピックと人名の関係

図 2 は『上井覚兼日記』内に出現する人名とそ

(c) Information Processing Society of Japan

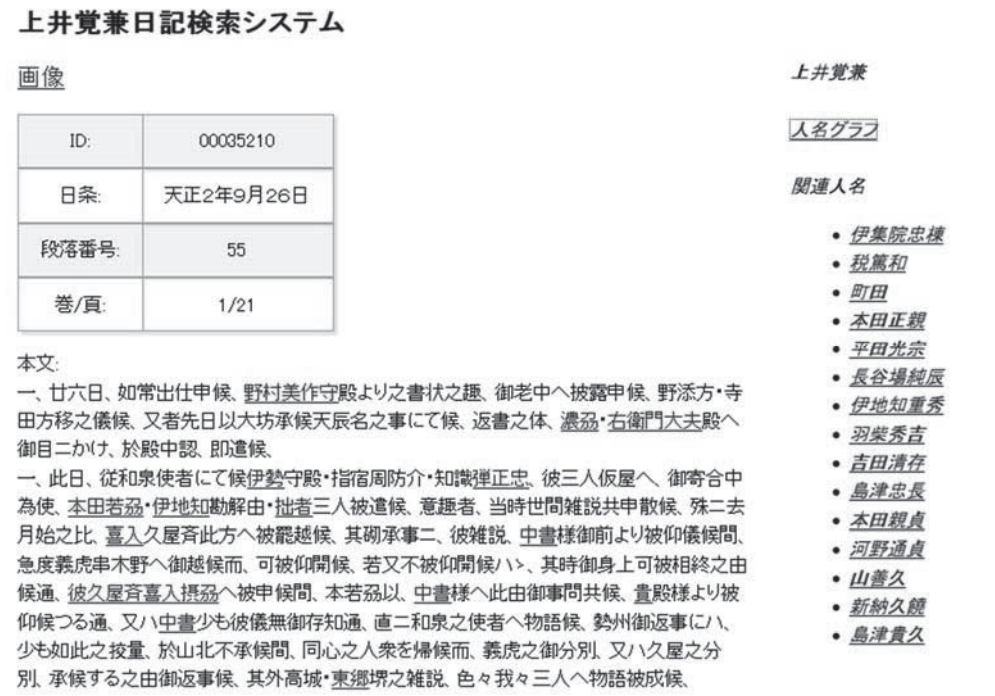

図８：検索結果詳細
Figure6: detail of search result

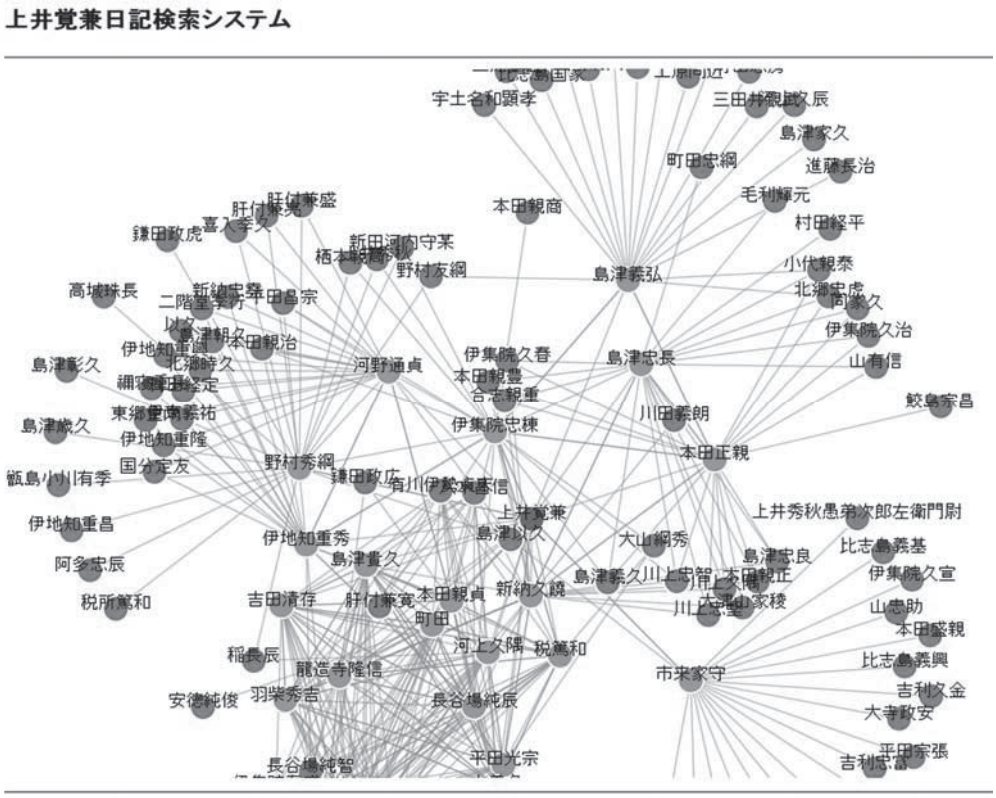

図９：人名関係の可視化
Figure9: visualization of relationship between personal names

の頻度を示す（上位 30 件）．この史料を著者である上井覚兼が最も出現頻度が高かった．史料内では“上井覚兼”もしくは“覚兼”ではなく“拙者”が出現する．上井覚兼が島津家の家老であったこ

(c) Information Processing Society of Japan

とに依るかと思われるが，島津 4 兄弟（島津義久，義弘，歳久，家久）のほか，伊集院忠棟や島津忠長などの島津家の中枢にいた人名の出現頻度が高かった．また，上井覚兼の家臣だった柏原有閑などの宮崎衆（『上井覚兼日記』において“宮崎衆”と記述がある．）や上井薫兼などの上井覚兼一門衆の頻度も高かった．宮崎衆の出現頻度が高かった理由は上井覚兼が日向国宮崎地頭・宮崎城主だったことであろうと思われる．

図 3 は検出した潜在トピックと各潜在トピックに関係する人名（出現頻度上位 20 件分）を示す．例えば，V1 は家老や肥後に関わる人名，V2 は大隅関係・肝付氏との交渉・抗争に関わる人名，V3 は龍造寺氏関係および肥前・筑前情勢に関わる人名，V4 は上井覚兼の一門衆・宮崎衆，V5・V8・V9・V10 は老中衆・島津一門衆を示すと考えられる．特に，V5 は島津歳久関係，V8 は当時の島津家の当主である島津義久関係，V9 は島津義弘関係，V10 は島津家久関係であろうと思われる．

４．２．人名間の関連性

図 4 は，島津 4 兄弟（島津義久・義弘・歳久・家久）との出現パターンが類似する人名の上位 15 名を示している．ここでは，各トピックに属する人名とその出現頻度をもとに，各人名間の距離を算出する．このとき人名－トピックのベクトルを作成し，これに次式で示すコサイン類似度を適用する．

$$\mathrm{score}(\mathrm{x},\mathrm{y}) = \frac{\sum_i x_i \cdot y_i}{\sqrt{\sum_i x_i^2} \cdot \sqrt{\sum_i y_i^2}} \qquad (2)$$

ここでx, yは人名ベクトルを示す．この値は各人名の出現パターンに対する類似度に相当する．

図 1 における V9 および V10 の結果を見ると，義久と義弘もしくは義弘と家久の出現パターンは類似するかと思われた．しかしながら，図 3 の結果から，各兄弟と出現パターンが類似する人名が必ずしも類似していないことがわかる．この結果より，島津 4 兄弟とのは必ずしもともに行動していないことがわかる．

島津義弘に近い出現パターンだったのは，兄弟の中では当時の島津家当主島津義久だったことがわかる．義弘は義久に代わって総大将として指揮を執ることが多かったと考えられるが，これが影響するかもしれない．

島津家久はその居城が日向国佐土原にあり，上井覚兼の居城は日向国宮崎である．この関係によるかもしれないが，家久と類似する人名は図 3 で示した V4 と比較すると上井覚兼の一門衆・宮崎衆が多く含まれていることがわかる．

４．３．トピックの時系列変化

図 5 に各トピックの時系列変化を示す．この図では，天正 2 年 8 月から天正 14 年 10 月までの期間を 3 ヶ月ごとに区切っており，期間ごとに出現するトピックの配合率を示している．図 5 より，トピックの配合率は一定ではなく，期間ごとに異なっており，時系列変化がある．

天正 2 年から天正 4 年の期間ではトピック V2 の割合が他のトピックに比べ大きいが，天正 10 年以降では反対にその割合は非常に小さくなっている．天正 2 年から 4 年の間は，天正 2 年はちょうど肝付氏・伊地知氏を下した時期である．この時期に大隅の情勢を多く書き残したのは，上井覚兼が鹿児島の奏者だったことに影響するかもしれない．一方で，天正 10 年以降では V2 の割いたことが要因の 1 つかもしれない．

天正 10 年以降ではトピック V4 の割合が，天正 2 年から 4 年に比べ大きい．前述のとおり，V4 に属する人名の多くは上井覚兼の一門衆・宮崎衆である．上井覚兼は，この日記には記述はないのだが，天正 8 年より宮崎城主となった．このことが影響していると考えられる．

図 6 はトピック V9 およびトピック V10 における時系列変化を示す．この図では，一部の期間について各トピックに属する人名を列挙した．図 3 からわかるように，V9 は島津義弘，V10 は島津家久に関係する人名が属しているが，両トピックとも，義弘・家久が属している．しかしながら，V9 には V1 や V2 に属する人名が多く，V10 には V4 に属する人名が多いことから，これらの特徴は大きく異なる．図 6 でもこの傾向が見受けられるが，両トピックとも天正 2 年から天正 4 年の期間と天正 10 年以降の期間では属する人名が異なることが多かった．例えば，天正 3 年 1 月 1 日－天正 3 年 3 月 29 日の期間において V9 に属する人名は，上井覚兼，上原尚近，平田歳宗，新納教久，島津義弘，本田親治，伊地知重増，伊地知重秀，伊集院忠棟などである．これに対し，天正 14 年 7 月 1 日－天正 14 年 9 月 28 日の期間では，上井覚兼，島津義久，島津義弘，上原尚近，名和顕孝，島津忠長，伊集院久治，城一要，宮内忠増などである．『上井覚兼日記』の日条において，共起する人名が明らかに異なっていることを示す．これは V10 も同様だった．

５．プロトタイプシステム

本節ではプロトタイピングした『上井覚兼日記』に対する検索システムについて述べる．

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し，検索結果の詳細表示を提

(c) Information Processing Society of Japan

示するとともに，テキスト内に出現する人名およびその人名に関連する人名を提示する機能を提供する．また，指定した人名と関連する人名をグラフィカルに提示する機能を有する．

本システムは，入力した検索語に応じて文字列探索を行い，図 7 のように検索結果を示す．この例は，”家久”を検索クエリとしたときの検索結果である．ここでの検索結果は古記録 FT における ID，和暦年月日，段落番号，本文，およびこの結果に対応する『大日本古記録　上井覚兼日記』の刊本画像へのリンクで構成される．本システムは KWIC 機能を有している．

検索結果のうち 1 件を選択すると，検索結果の詳細を表示することができる．これを図 8 に示す．この例は「天正 2 年 9 月 26 日条」を示しており，ID，日条，段落番号，『大日本古記録 上井覚兼日記』における巻・ページ，およびテキストを提示する．該当するテキスト内に出現する人名をマークアップしている．人名の抽出は 3.2 節で説明した手法に従った．本システムでは（2）式により求められた score(x,y)が 0.1 以上であれば関連する人名として提示することにした．また，この値に応じて提示する人名をソートしている．マークアップされた人名に対しマウスオーバーすると，ツールチップにより統制された人名が提示される．人名をクリックすると，関連する人名が提示される．関連する人名は 4.2 節で述べた手法に従った．

図 8 から”人名グラフ”をクリックすると，関連する人名をグラフィカルに提示する．これを図 9 に示す．ここでは指定した人名から 2 ホップ先まで（関連する人名に対する関連する人名まで）を提示する無向グラフである．起点となる人名は青で，1 ホップ先はオレンジで，2 ホップ先は緑でノードを色分けしており，関連する人名間はエッジでつないでいる．このグラフは Force-Directed Graph Drawing と呼ばれており，このプロトタイプシステムでは D3.js[8]を用いて表示している．

6．おわりに

本論文では，天正期古記録『上井覚兼日記』に対して，人名抽出手法，LDA を用いた潜在トピックの検出，および人名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『上井覚兼日記』検索システムの概要，テキスト検索機能，関連人名の提示機能について述べた．

LDA は教師なし学習であるため学習データを必要としない．これにより，『上井覚兼日記』以外の史料にも適用可能であるため，適用範囲は広い．また，共起性に基づく潜在トピックの検出を行うことができるため，主観的ではなく，客観的な人名間の関係を示すことができる．今後，日本史史料に関するテキストは，今以上に膨大に作成されるようになると思われる．そのような状況において本手法は非常に有用であると考えられる．また，どのような要因で関連するのか，ということをテキストまで遡って解析できることも本手法の利点であろう．4 節で述べたとおり，史料全体に対する解析や時系列変化の解析は容易に行うことができる．

本研究における人名の抽出手法はテキストの記述方法に依るところが大きく汎用性はないと考えている．今後は人名抽出手法の洗練に努めたい．また，1 史料だけでなく，地名や他の用語を抽出し関係付けていく，さらに同時代の同地方の史料を加味するなど，日本史学的分析を推し進める手法について考慮していく予定である．

謝辞

本研究の成果の一部は，日本学術振興会科学研究費若手研究（B）（26730167）および基盤研究（A）（70143534）の助成を受けたものによる．

参考文献

1) 東京大学文学部・大学院人文社会系研究科: “日本史学研究室”，入手先〈http://www.l.u-tokyo.ac.jp/laboratory/database/8.html?department=2〉（参照 2014-11-07）.
2) 国立歴史民俗博物館 : “データベースれきはく”，入手先〈https://www.rekihaku.ac.jp/education_research/gallery/database/index.html〉（参照 2014-11-07）.
3) 東京大学史料編纂所 : “データベース検索”，入手先〈http://wwwap.hi.u-tokyo.ac.jp/ships/〉（参照 2014-11-07）.
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) 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.
6) 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.
7) 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.
8) Mike Bostock : “D3.js - Data-Driven Documents”, 入手先〈http://d3js.org/〉（参照 2014-11-07）.

(c) Information Processing Society of Japan