様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　６　月　　１　日現在

機関番号：　１２６０１
研究種目：　若手研究(B)
研究期間：　2014～2017
課題番号：　２６７３０１６７

研究課題名（和文）テキストマイニング分析による史学的知識抽出に関する研究

研究課題名（英文）A study on historical knowledge extraction by text mining analysis

研究代表者
山田　太造（YAMADA, Taizo）

東京大学・史料編纂所・助教

研究者番号：７０４１３９３７

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：本研究では日本史学研究における研究過程支援のため，史料目録・テキスト等から史学的知識を抽出し，知識間・知識－史料間・史料間などの関係を明確にしながら，内在する史学的知識・暗黙知を外在化する研究を行うため，特に（1）史学的知識の抽出・蓄積，（2）史学的知識間，史料学的知識－史料間，史料間の関連性の検出，（3）知識表現・関連表現とそれらを用いた検索手法の確立を目指した.

研究成果の概要（英文）：In the study, in order to support the research process in Japanese historical research, we externalize the inherent historical knowledge and tacit knowledge by extracting historical knowledge from historical catalog and text and clarifying relations such as inter-knowledge / knowledge-historical materials / historical materials. Particularly, we tackled (1) extraction and accumulation of historical knowledge, (2) relation detection of inter-historical knowledge, historical knowledge-historical materials and between historical materials, (3) knowledge representation and related expressions and search using them.

研究分野：　データ工学

キーワード：　歴史情報　日本史史料　テキストマイニング　セマンティックウェブ　情報検索　メタデータ

１．研究開始当初の背景

日本史学研究を展開する場合，まず研究素材となる日本史史料を調査・蒐集，読解，史料批判を行う必要がある．調査・蒐集では，国内外に点在する日本史に関係する史料について，史料名・所蔵情報・時間情報・形態などのメタデータを集約し，マイクロフィルム撮影・デジタル撮影などによる複写を行い，デジタル画像化を行う．次に，採訪により蒐集した史料をもとに，翻刻・索引付け・標出作成などの読解を行う．また歴史研究を進める上で実証可能であるかを判断する史料批判も行う．これらの過程を経ることで論点を引き出し，歴史モデルを構成する．各過程で生成される情報は膨大であるが，多様な史料をもとに実証していく必要がある．これまで，東京大学史料編纂所（以下，史料編纂所）歴史情報データベース（以下，SHIPS DB），国文学研究資料館日本古典籍総合目録データベース，国立国会図書館サーチ（以下，NDL Search）などに代表されるデータベース検索サービスにより，日本史史料の目録・画像・テキストを検索することができるようになり，調査・蒐集のタスクを軽減することができるようになった．しかしながら，読解や史料批判などを行った結果を用いた検索・利活用を可能とする方法論，およびシステム化はまだ確立していない．特に，関連する人物・空間・時間などの史学的知識を用いて検索し，史料を関連付ける方法論について，量的に膨大に扱うことを念頭に置いた方法論については確立していない．日本中世史，特に戦国時代における史料研究では，扱う史料の点数は膨大であるため，その時代を全体的に議論すること自体が困難である．そのため，特定の家・寺社・地域を扱うなど，ドメインを絞り込むことで研究を進展している．しかしながら，日本史として展開する上では全域における歴史像を見出すことは不可避である．そのため，日本史学研究における研究過程，特に読解・史料批判を支援するために，（1）史学的知識の抽出・蓄積，（2）史学的知識間，史学的知識-史料間，史料間の関連性の検出，（3）知識表現・関連表現とそれらを用いた検索手法の確立，を行い，内在する研究的知識・暗黙知の外在化を目指す．

２．研究の目的

前述の研究の学術的背景に基づき，本研究では以下について明らかにする．本研究では，SHIPS DB における各データを素材とすることを想定する．SHIPS DB には史料目録・画像，フルテキスト，索引（人名・地名・事項）などを検索対象としてサービスしている．

（1）史学的知識の抽出・蓄積

文献史料は，古文書，古記録，聖教の 3 つに分類することができる．古文書の場合，記述された時間，さらに差出・宛所の人物情報，記述された文面から，人物・空間情報，事件・事項などが抽出できると考えられる．また古記録の場合，記述した人物，各日の事件・事項・人物・空間情報などが，聖教の場合であれば，奥書より人物・時間情報などが抽出できると考えられる．これらの史学的知識は，これまでは人手で抽出されてきた．しかしながら，日本中世史で扱う史料は膨大な点数であり，調査・収集により新史料が発見されることも稀ではないため，人手での抽出には限界がある．そこで，自動的に抽出する手法を検討する．史料テキストからの抽出に加え，史料目録や索引などの情報を組み合わせることで，より洗練された史学的知識の抽出を行う．

また，「史料学研究支援のためのアノテーション管理基盤に関する研究」（若手研究（B），2009-2011 年度）において実施したアノテーションを利活用した研究的知見の情報基盤の成果を取り入れつつ，実施していく．

（2）史学的知識間・史学的知識-史料間・史料間の関連性の検出

ベクトル空間モデルやトピックモデルの 1 つである LDA（Latent Dirichlet Allocation）などの機械学習手法を適用することで，史学的知識間・史学的知識-史料間・史料間の関連性の検出を行う．特に LDA では潜在する話題に基づき分類を行うことが可能であるため，文字列だけでの共起関係から意味的な要因の共起関係を扱うことが可能となる．意味的な要因の共起関係を分析することで，史学的研究における暗黙知・暗黙的理解を明示できると期待している．

（1）により抽出・蓄積した知識をもとに，各史料を分析し，ベクトル空間モデルや LDA により，史学的知識と史料の関係を明示する．また，史学的知識の共起関係を検出することで知識間の明示化を図る．さらに，史料間の関係を知識間関係・知識-史料間関係，および，史料目録における史料間の関係を取り入れながら，明示化を進める．

（3）知識表現・関連表現とそれらを用いた検索手法の確立

（1）により検出・蓄積された知識を機械可読も可能な形式で表現していく．このとき RDF や Linked Data などの semantic web 技術を考慮して進める．また，（2）における関連性についての表現も同様に進めていく．蓄積した知識・関連を用いた検索手法も考慮していく．例えば，同一の話題である史料や，空間・時間などの要素も加味した検索を実現していく．さらに，各種史料と関連する史学的知識や史料も合わせて提示する手法についても検討を進める．

３．研究の方法

本研究では日本史学研究における研究過程支援のため，史料目録・テキストから史学的知識を抽出し，知識間・知識-史料間・史料間などの関係を明確にしながら，内在する史学的知識・暗黙知を外在化する研究を行うため，以下の問題を解決する．

（1）史学的知識の抽出・蓄積
本研究では，SHIPSDB における各データを素材とすることを想定する．SHIPSDB における「所蔵史料目録 DB」や「日本古文書ユニオンカタログ」などでは史料目録，「古文書フルテキスト DB」や「古記録フルテキスト DB」などでは史料テキスト，「大日本史料総合 DB」や「中世記録人名索引 DB」などでは索引（人名・地名・事項など）を提供している．史学的知識の抽出においては，史料テキストから抽出するとともに，史料目録や索引情報も用いる．史料テキストからの抽出では，前述した先行研究により，日本南北朝期史料のテキストを対象とした単語分割が可能であるため，この手法を用い，さらなる精度向上を図る．また SHIPSDB における索引情報を用いるなどにより，史学的知識に関するタギングを行う．未知である場合については改めて検討するが，「（2）史学的知識間，史学的知識-史料間，史料間の関連性の検出」で述べる方法を用いることで，自動的にタギングできる方法について検討し，最適な手法を探求する．また，知識の蓄積方法についても検討する．ここでは「（3）知識表現・関連表現とそれらを用いた検索手法の確立」の方法と連動しながら，最適な手法を探求する．
（2）史学的知識間，史学的知識-史料間，史料間の関連性の検出
抽出した知識を蓄積することにより，知識間や知識-史料間での新たな関係を見出す．このとき，LDA などのトピックモデルなどを利用することで客観的な関係を見出す．ここでの方法は，基本的には教師無し学習的方法を応用することで実現していく．また，時空間情報などを史学的知識と関係づけていくことで，いつ・どこで・なにに関する知識であるかが明確になっていき，時系列変化・空間的変化を追跡できるようになる．さらに SHIPS DB 内に知識，および，知識間関係が格納してあれば，それも用いて半教師あり学習的方法により，新たな関係もしくは見いだしてきた関係に対するサポートなど，洗練された知識となりうる．洗練された史学的知識をもとに史料間を関係づけていく．ここでの史料間関係は，これまでの史料目録や単語共起をベースとする関係だけでなく，意味的・潜在的な要因に基づくものであろうと考えられる．
（3）知識表現・関連表現とそれらを用いた検索手法の確立
知識や知識間関連などの表現においては，RDF や Linked Data に代表される sematic web の適用可能性の検討を進める．表現方法は，抽出される史学的知識や各種関連に依存すると考えられるが，汎用的に利用可能な表現形式をデザインしていく．また，オープンアクセス可能な提供方法を模索する．ここでのオープンアクセスの対象は，抽出した知識，知識間関連，および，典拠となる史料を想定している．また，SHIPS DB における各種史料検索サービスにおいて，これらの知識や各種関連を用いた検索手法について検討し，最適な方法を模索する．ここで，デザインされた知識および各種関連を用いた検索を実現すべく，検索インターフェースのプロトタイピングを進めていく．

４．研究成果
平成 26 年度は，史学研究者が研究過程で蓄積している史学的知識，特に天正記の史料を用いて人名・時間に着目し，調査・分析を進め，分析の結果に基づき，知識および各種関連の表現方法，さらに蓄積方法を提案するため，上記の（1）および（3）についてデザインし，プロトタイピングを進めた．これにより，ある人物を時系列に追跡し，人名間の関係を提示するなどの日本史における人名に関するテキストマイニングの実現を進めた．最初の段階であるため，人名・時間に特化してシステムのプロトタイピングしテキストマイニングの結果を分析した．しかしながら，他の知識については扱っていないため，知識表現・各種関連のデザインや蓄積のためのプロトタイプシステムの完成度は低く，上記の（3）における検索手法としては，十分な機能を持ち合わせていないと考えられる．この後，史学研究者から初期段階でのプロトタイプの利用を通じた率直な意見を収集し，それを分析することでよりよいシステムを構築していくことができると考えている．
平成 27 年度は，（1）についてさらに深化させ，天正期の古記録テキストから精度の高い史学的知識の抽出を行った．（2）について検討を進める上で，対象史料の拡大は必須であると考え，尊卑分脈や柳営補任のような家系図や任免記録も対象として史学的知識の蓄積を行い，これらの結果をもとに，（3）についての対象範囲を拡大するための改善を行った．さらに（3）について，RDF を用いた知識表現・関連表現方法を検討した．
平成 28 年度は，（1）について天正期以外へ適用するなどによりさらに深化させた．さらに，これまで進めてきたセマンティックウェブ技術，具体的には RDF を用いた知識表現およびその蓄積方法を適用し，蓄積してきた人名の関連を抽出してきた．（3）については蓄積した史学的知識に対するセマンティックウェブ技術を用いた検索方式，具体的には SPARQL を用いた検索手法を提案し，さらにそれをプロトタイピングしたウェブシステム上にて適用するなど，さらなる深化に努めた．（2）について，時代・性格の異なる史料群を対象とすることで知識等の関連の検出するため，系図類や任免記録等を，史学研究者とともに性格や史学研究における利用等について洗い出しを行い，史料−史学的知識間の関連について検討を進めた．これまでの研究成果を人文科学とコンピュータシンポジウムじんもんこん 2016 や国際会議 Digital Humanities 2016 などで発表した．

平成 29 年度は，人名検索システムをさらに深化させ，人名とその典拠史料を示し，さらに時系列に分析できる機能を設けた．また空間的分析手法についても提案した．これまでの研究成果を人文科学とコンピュータシンポジウムじんもんこん 2017 や国際会議 Digital Humanities 2017 などで発表した．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 26 件）

① 山田太造、日本史史料にもオープン化が到来-歴史学研究波動変貌していく?-、歴博 207 号、査読無、207、2018、6
② Taizo Yamada、Detection of topics from newspaper and its analysis of temporal variations in regions 、 proc. of PNC2017、査読有、2017、2017、44-49
③ Taizo Yamada, Satoshi Inoue 、Collecting the Name of a Historical Person from Related Historical Material、proc. of DH2017、査読有、2017、2017、807-808
④ 山田太造、新聞記事に対するトピックモデルの適用とトピックの時系列変化に関する考察、研究報告人文科学とコンピュータ(CH)、査読無、2017-CH-115、2017、1-5
⑤ 山田太造、人文科学と情報学の学際領域における課題、研究報告人文科学とコンピュータ(CH)、査読無、2017-CH-114、2017、1-3
⑥ 山田太造、企画セッション「構築したシステムのゆくえ」の概要、研究報告人文科学とコンピュータ(CH)、査読無、2017-CH-113、2017、1-3
⑦ 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、査読有、2016、2016、159-164
⑧ 山田太造、トピックモデルを用いた日本史史料テキストの分析手法、建築雑誌 2016 年 11 月号、査読無、1690、2016、36-37
⑨ Taizo YAMADA 、Classification and Representation of Scenes in Field Note by Spatiotemporal Characteristics Using Linked Data、proceedings of PNC 2016 Annual Conference and Joint Meetings、査読有、2016、2016、-
⑩ 山田太造、史資料からの地理空間情報の収集と管理に関する考察、研究報告人文科学とコンピュータ(CH)、査読無、2016-CH-111、2016、1-6
⑪ Taizo Yamada 、Satoshi Inoue 、A Management of Personal Name with Alternate Name and its Searching for Japanese Historical Study 、Proceedings of Digital Humanities 2016、査読有、2016、2016、716-717
⑫ 山田太造, 井上聡、前近代日本史史料に関わる人名情報の収集・蓄積に関する考察、研究報告人文科学とコンピュータ(CH)、査読無、2016-CH-109、2016、1-4
⑬ 柳澤雅之, 高田百合奈, 山田太造、地域情報学の読み解き-発見のツールとしての時空間表示とテキスト分析、地域研究、査読有、16、2016、267-291
⑭ 山田太造、東京大学史料編纂所の編纂とその業務にともなうデータベース、「資料がつなぐ大学と博物館―「研究循環アクセスモデル」の構築にむけて ―」予稿集、査読無、-、2016、52-55
⑮ Taizo YAMADA 、Extraction and Management of Spatiotemporal Term from Field Notes and Data Structuring for its Sharing in Area Studies、Proceedings of PNC2015 Annual Conference and Joint Meetings、査読有、 2015、2015、-
⑯ Taizo YAMADA, Satoshi Inoue、Detection of People Relationship Using Topic Model from Diaries in Medieval Period of Japan、Proceedings of DH2015、査読有、2015、2015、-
⑰ 山田太造、フィールドノートに記述された場面を特徴づける-語彙による知識処理-、特集第 20 回情報知識学フォーラム「地域情報学における知識情報基盤の構 築と活用」、査読無、25、2015、315-324
⑱ 山田太造、地域研究史資料を対象とした時空間的特徴の抽出と場面の構造化、第 14 回情報科学技術フォーラム講演論文集、査読無、14、2015、409-410
⑲ 山田太造、地域研究資料と対象とした時空間情報に着目したデータの構造化、人文科学とコンピュータ研究会報告、査読無、2015-CH-105、2015、1-6
⑳ 山田太造、日本史史料を対象としたテキスト構造化と読解支援、東京大学史料編纂所[共同研究拠点と歴史情報]シンポジウム「資料情報の新たな発信」 予稿集、査読無、2015、20-24
21 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、査読有、2014、2014、131-138
22 高田百合奈, 渡邉英徳, 柳澤雅之, 山田太造、位置情報とトピックモデルに基づくフィールドノートのビジュアライズ手法、じんもんこん 2014 論文集、査読有、2014、2014、57-62
23 清野陽一, 山田太造, 高田智和, 古瀬蔵、人文科学データベースからの人名一覧表示システムの構築、人文科学とコンピュータ研究会報告、査読無、2014-CH-103、2014、1-6

24 Taizo Yamada、Place Name Extraction from Field Notes Based on Text Analysis for Area Studies、Proceedings of PNC2014 Annual Conference and Joint Meetings、査読有、2014、2014、55
25 Taizo YAMADA, Satoshi INOUE、A Text Encoding Support System for Pre-modern Japanese Historical Materials 、 Proceedings of Digital Humanities 2014、査読有、2014、2014、558-559
26 山田太造、前近代日本史史料をベースとしたテキストデータベースの特徴と課題、日本語学臨時増刊号、査読無、2014年11月号、2014、29-41

〔学会発表〕(計 30 件)

① Taizo YAMADA、Flow and Utilization of Japanese Historical Data in the Historiographical Institute 、International Symposium "DIGITAL HUMANITIES AND DATABASES"(招待講演)(国際学会)、2018年3月、上智大学比較文化研究所(東京都千代田区)
② 山田太造、追加した新聞記事に対するトピック検出、H-GIS研究会、2018年1月28日、函館コミュニティプラザGスクエア(北海道函館市)
③ 山田太造、前近代日本史史料における人物関係とその時空間変化:天正期古記録『上井覚兼日記』を例に、じんもんこん2017、2017年12月9日、大阪市立大学学術情報総合センター本館(大阪府大阪市)
④ Taizo YAMADA、Detection of topics from newspaper and its analysis of temporal variations in regions、PNC2017(国際学会)、2017年11月6日、台南(台湾)
⑤ 山田太造、トピックモデルを用いた史資料の分析手法、数理地理モデリング研究会(招待講演)、2017年11月1日、京都大学数理解析研究所(京都府京都市)
⑥ 山田太造、古文書データの次の"切り口"を探す -古文書をさらに利活用していくために-、東寺百合文書データミーティング(招待講演)、2017年10月26日、京都府立京都学・歴彩館(京都府京都市)
⑦ Taizo YAMADA 、 Japanese History Research by Historiographical Institute the University of Tokyo and its Contribution、14th International Conference on Digital Preservation (iPRES2017)(国際学会)、2017年9月26日、京都大学国際科学イノベーション棟(京都府京都市)
⑧ 山田太造, 谷昭佳, 保谷徹、東京大学史料編纂所による前近代日本史史料の調査に基づく史料画像のデジタル化とその保存、14th International Conference on Digital Preservation (iPRES2017)(国際学会)、2017年9月25日、京都大学国際科学イノベーション棟(京都府京都市)
⑨ Taizo YAMADA、Collecting the Name of a Historical Person from Related Historical Material 、 Digital Humanities 2017(国際学会)、2017年8月9日、モントリオール(カナダ)
⑩ 山田太造、新聞記事に対するトピックモデルの適用とトピックの時系列変化に関する考察、第115回人文科学とコンピュータ研究会発表会、2017年8月4日、東京大学史料編纂所(東京都文京区)
⑪ 山田太造、史料編纂所歴史情報処理システムの今と新たな日本史情報の活用、東京大学史料編纂所公開研究会、2017年6月2日、東京大学史料編纂所(東京都文京区)
⑫ 山田太造、人文科学と情報学の学際領域における課題、第114回人文科学とコンピュータ研究会発表会、2017年5月13日、龍谷大学アバンティ響都ホール(京都府京都市)
⑬ 山田太造、電子くずし字字典データベースにおける現状と展望、第2回CODHセミナーくずし字チャレンジ−機械の認識と人間の 翻刻の未来−(招待講演)、2017年02月10日- 2017年02月10日、国立情報学研究所(東京都千代田区)
⑭ 山田太造、史資料と考古資料を利用していく環境、第113回人文科学とコンピュータ研究会発表会、2017年02月04日- 2017年02月04日、愛知工業大学本山キャンパス(愛知県名古屋市)
⑮ 山田太造、前近代日本史史料から人名を集める、人文科学とコンピュータシンポジウム「じんもんこん2016」、2016年12月11日- 2016年12月11日、国立国語研究所(東京都立川市)
⑯ 山田太造、東京大学史料編纂所における日本史史料の収集とその管理、東アジア日本研究者協議会第一回国際学術大会(国際学会)、2016年11月30日- 2016年11月30日、仁川(韓国)
⑰ Taizo YAMADA、An example of Collection and Digitalization of Materials Related to Japanese History、Workshop on the Academic Asset Preservations and Sharing in Southeast Asia(招待講演)(国際学会)、2016年11月20日- 2016年11月20日、バンコク(タイ)
⑱ Taizo YAMADA 、 Text structure of Japanese history historical materials and effort for applying TEI in Historiographical Institute of the University of Tokyo 、 The 1st International Workshop on Models of Japanese Texts and TEI(招待講演)(国際学会)、2016年08月31日- 2016年08月31日、東京大学経済学研究科交流棟

(東京都文京区)

⑲ 山田太造、東京大学史料編纂所における史料情報とその管理、第2回ナショナルデジタルアーカイブ研究会(招待講演)、2016年08月27日- 2016年08月27日、国立国会図書館東京本館(東京都千代田区)

⑳ Taizo YAMADA、Classification and Representation of Scenes in Field Note by Spatiotemporal Characteristics Using Linked Data、PNC 2016 Annual Conference and Joint Meetings(国際学会)、2016年08月17日- 2016年08月17日、ロサンゼルス(アメリカ合衆国)

21 山田太造、史資料からの地理空間情報の収集と管理に関する考察、第111回人文科学とコンピュータ研究会発表会、2016年07月30日- 2016年07月30日、福江文化会館(長崎県五島市)

22 Taizo YAMADA、A Management of Personal Name with Alternate Name and its Searching for Japanese Historical Study、Digital Humanities 2016(国際学会)、2016年07月15日- 2016年07月15日、クラクフ(ポーランド)

23 山田太造、テキストデータを使うとどのようにフィールドが分類できるか?、日本人口学会開催地域部会 2015 年度研究会(招待講演)、2016年03月05日- 2016年03月05日、総合地球環境学研究所(京都府京都市)

24 山田太造、地域研究史資料に対するテキストマイニング適用の試み、H-GIS研究会、2016年02月20日- 2016年02月20日、熊本県立大学(熊本県熊本市)

25 Taizo YAMADA、Analysis of Archaeological Information Using Topic Model Technologies、International Workshop on Application of Science and Technology for Cultural Studies(IWASTCS2015)(招待講演)(国際学会)、2015年11月13日- 2015年11月13日、Maha Chakri Sirindhorn Anthropology Centre (SAC), Bangkok, Thailand

26 山田太造、フィールドノートにおける場面分類とwebサービスを用いた地名抽出、H-GIS研究会、2015年09月12日- 2015年09月12日、京都大学地域研究統合情報センター(京都府京都市)

27 山田太造、史料編纂所における人名・地名に関するデータについて、H-GIS研究会、2015年05月30日- 2015年05月30日、京都大学地域研究統合情報センター(京都府京都市)

28 山田太造、歴史学の情報?、イベント企画:IPSJ-ONE, 情報処理学会第77回全国大会(招待講演)、2015年03月17日-2015年03月17日、京都大学百周年時計台記念館百周年記念ホール

29 山田太造、地域研究データにおけるトピックの検出と時空間変化に関する研究、H-GIS研究会(招待講演)、2014年10月17日-2014年10月17日、京都大学地域研究統合情報センター

30 Taizo Yamada、Text mining for Historical Documents & RDF and Linked Open Data-、Pre-Symposium of Kyoto University ASEAN Center (Bangkok Office) Opening Ceremony and Commemorative Symposium(招待講演)、2014年06月27日-2014年06月27日、Grand ballroom, 4th Floor Grand Millennium Sukhumvit, Bangkok, Thailand

〔図書〕(計 3 件)

① 山田太造、東京大学史料編纂所の編纂とその事業にともなうデータベース、吉川弘文館、〈総合資料学〉の挑戦、2017、180(98-113)

② 山田太造、ガラス乾板に関するデータはどこに向かうのか、勉誠出版、文化財としてのガラス乾板、2017、262(180-183)

③ 山田太造、文字データベース連携の課題、勉誠出版、漢字字体史研究 二、2016、432(395-419)

６．研究組織

(1)研究代表者

山田太造（YAMADA, TAIZO）
東京大学・史料編纂所・助教
研究者番号：７０４１３９３７