Vol.2013-CH-97 No.6
2013/1/25

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-

山田　太造[1]

概要：昨今，日本史史料に関して目録や画像だけでなく本文もデジタル化が進み，公開されつつある．本研究では，日本の南北朝期における史料を対象に，本文を用いて類似する史料を収集し，提示する手法について紹介する．

キーワード：日本史史料，テキストマイニング，機械学習，n-gram 言語モデル，ノンパラメトリックベイズ

Study on a Method of Collecting Similar Japanese Historical Materials in Nanboku-cho Period of Japan

Yamada Taizo[1]

Abstract: Recently, in Japanese historical material, the amount of encoded texts has been grown up, and encoded texts have been published by web-based systems. In the paper, using similarity between text of Japanese historical materials which are written in Nanboku-cho period of Japan, we introduce a method to collect and represent the materials.

Keywords: Japanese historical material, text mining, machine learning, n-gram language model, nonparametric bayes

1. はじめに

昨今，日本史史料において，目録や画像だけでなく，本文もデジタル化が進み，公開されつつある．日本史研究に関連する全文データベースとして，東京大学史料編纂所歴史情報処理システム（SHIPS：Shiryohensanjo Historical Information Processing System）における全文データベース[*1]，国文学研究資料館『吾妻鑑』データベース[*2]，国立歴史民俗博物館記録類全文データベース『兼顕卿記』[*3] などがある．これらのシステムには，本文に対し，ユーザが入力した文字列が完全一致した箇所を，その前後の文字列とともに表示する kwic（keyword in context）機能や，ヒットした史料の目録や画像を提示する機能を有しているものもある．さらに，日付，函・巻，管理番号のような史料目録情報に従ってソートすることもできる．

全文データベースは目録型のデータベースに比べ，当然のことながら，文字数が圧倒的に多くなる．また，今後より一層本文のデジタル化が進むと予想される．そのため，大量のテキストを効率的に分析しうる仕組みが求められるようになる，と考えている．そのためにはクエリとの関連性に基づく検索結果のソート，関連するコンテンツの提示，提示したコンテンツの可視化などの仕組みが必要であろう．しかしながら，そのような機能を提供するシステムはほとんどない．

本稿では，日本史史料テキストを用いて，クエリとの類似性に基づくランキング，関連する史料の提示，検索結果の可視化の機能を実現し，それらの機能がもたらす日本史史料分析への寄与を検証する．クエリに合致した史料をラン

[1] 人間文化研究機構
National Institutes for the Humanities

[*1] http://wwwap.hi.u-tokyo.ac.jp/ships/shipscontroller
[*2] http://base1.nijl.ac.jp/~anthologyfulltext/
[*3] http://www.rekihaku.ac.jp/doc/t-db-index.html

text:	御教書案師直師泰誅伐事早馳参御方可致軍忠之状如件観応元年十一月三日御判島津左京進入道殿
result:	御教書 | 案 | 師直師 | 泰誅伐事 | 早馳 | 参御方 | 可致 | 軍忠之 | 状如件 | 観応元年十一月三 | 日 | 御判 | 島津 | 左京進 | 入道殿
correct?:	御教書 | 案 | 師直 | 師泰 | 誅伐事 | 早馳参 | 御方 | 可致 | 軍忠之状 | 如件 | 観応 | 元年 | 十一月 | 三日 | 御判 | 島津 | 左京進 | 入道 | 殿

図 1 古文書テキストの例（『島津家文書 足利直義御教書案（切紙））』

キングするためには，個々の合致した史料におけるクエリに関するスコアを検索エンジンが計算しえる仕組みが必要である．また，史料間の類似性を計算するために，史料間に関するスコアの計算も必要である．2 節では，クエリ－史料間，もしくは史料間の類似性を計算するため，ベクトル空間モデル（vector space model）[2] を導入する．ここでは，各史料を重み付け空間のベクトルとしてみなすことにより，クエリ－史料間もしくは史料間のスコアを計算する．ベクトル空間モデルでは各史料はベクトルとして表現するため，ベクトルの要素となりうる用語を抽出する必要がある．そこで 3 節において日本史史料のテキストから用語を抽出する手法について述べる．ここでは，NPYLM（Nested Pitman-Yor Language Model）[4] というノンパラメトリックベイズの手法に基づく手法を用い，史料テキストにおける文を単語分割し，その結果を用語として用いた．4 節では，構築したプロトタイプシステムの概要について示す．また，時系列的に関連史料を可視化できる機能についても示す．5 節では，日本史史料の分析への寄与と今後の展望について述べる．

2. 史料間の類似性

クエリ－史料間もしくは史料間の類似性に基づき関連する史料をソートするために，ベクトル空間モデル [2] を導入する．

ベクトル空間モデルでは，各史料を重み付けベクトルとして表現する．そのため，ベクトルの各要素は史料テキストから抽出した各用語の重みとなる．ある史料 x における用語 i の重み $weight(x)_i$ を次式に示す tf-idf 重み付けにより計算する．

$$weight(x)_i = tf(x)_i \cdot \left(\log \frac{N}{df(i)+1} \right) \quad (1)$$

ここで，$tf(x)_i$ は史料 x における用語 i の出現頻度，$df(i)$ は用語 i を含む史料の個数，N は史料数を示す．tf-idf 重み付けは i が少数の史料で多く出現するとその重みを大きくし，反対に，多くの史料で出現すると重みを小さくする．そのため，重みの大きな用語はその史料のテキストの性質を特徴付けると考えられる．

史料間の類似性を量化するため，次式で求める 2 つのベクトル x と y のコサイン類似度（cosine similarity）を計算する．

$$sim(x,y) = \frac{\sum_i weight(x)_i \cdot weight(y)_i}{\sqrt{\sum_i weight(x)_i^2} \cdot \sqrt{\sum_i weight(y)_i^2}} \quad (2)$$

本研究では，この値を史料間の類似性を示すスコア（類似度）として用いることにした．

$$score(x,y) = sim(x,y) \quad (3)$$

この $score(x,y)$ を用いることで史料間の類似度に応じたランキングを実現することができる．

クエリ－史料間の類似度を計算するためには，まず検索時にクエリのベクトルを作成し，次に式（2）を計算する．この値に応じてランキングすることで，検索結果をクエリとの類似度の応じてソートすることが可能となる．

3. 用語抽出

ベクトル空間モデルにおいて，式（1）を計算するためには，史料テキストから用語を抽出する必要がある．**図 1** に古文書テキストの例（『島津家文書 1 足利直義御教書案（切紙））』（文書番号 562））を示す．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[*4]や mecab[*5]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．[5] 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，残念ながら古文書・古記録への適用はまだ困難な状態にある．また，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．

図 1 をみると，出現する品詞としては名詞が多く，動詞等他の品詞は少ない．これは他の古文書・古記録でも同様の傾向にある．そこで，本研究では，単語分割した結果を

[*4] http://chasen-legacy.sourceforge.jp/
[*5] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

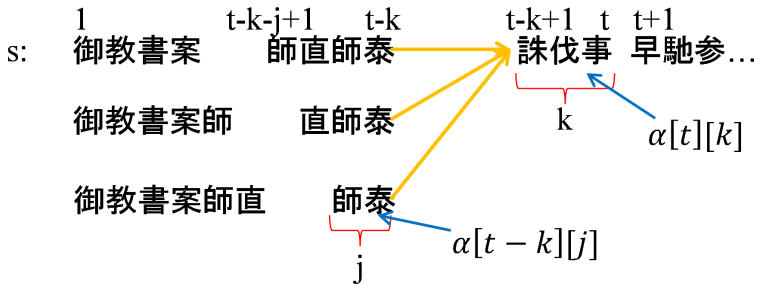

図 2 Forward filtering

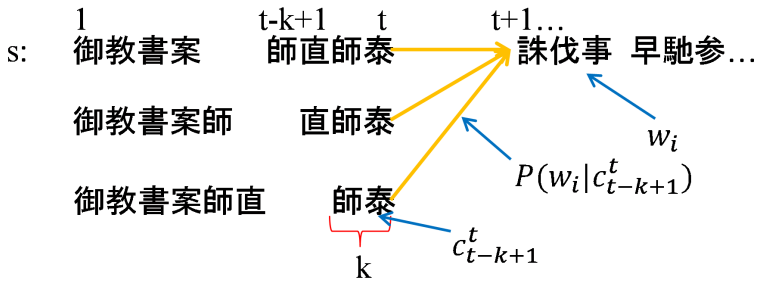

図 3 Backword Sampling

用語として扱うことにした．

日本史史料テキストに対する単語分割の手法自体は，知りうる限りにおいては，文献 [4] 以外皆無である．文献 [4] の手法は，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単語らしさを計算し，推定していく．この手法を用いて単語分割した結果は図 1 に示したとおりである．この結果より，思った通りとはいかないものの，正解と思われる単語分割と遜色ない結果が得られていると考えている．

この単語分割手法では，最適は単語分割を推定するために，Forward filtering-Backward sampling 法を用いる．ここでは文字列から文字列への遷移確率を動的計画法により求めていくが，この遷移確率を NPYLM という n-gram 言語モデルにより算出する．また，この単語分割手法は Gibbs Sampler[1] と呼ばれる動的計画法を用いてサンプリングしていく．このとき，すべての文に対し Forward filtering-Backward sampling 法を適用し，これから得られた単語分割をもとに言語モデル，および単語分割のためのパラメータを更新する．また，各種パラメータが収束もしくは，その変動が小さくなるまで繰り返し行う．

この節では，これ以降，Forward filtering-Backward sampling および NPYLM について簡単に述べる．詳細は文献 [4] を参考にされたい．

3.1 Forward filtering-Backward sampling

ここでは，単語分割 $w(s)$ を推定するための Forward filtering-Backward sampling 法について述べる．この手法は Forward filtering と Backward sampling の 2 つのフェーズがある．

3.1.1 Forward filtering

文字列から文字列への遷移確率 $\alpha[t][k]$ をバイグラムで表現する．**図 2** に示すように，$\alpha[t][k]$ は，文字列 s の部分文字列 $c_1, \ldots, c_t$ から最後の k 文字を単語として生成した場合の確率を示す．これは次式で計算する．

$$\alpha[t][k] = \sum_{i=1}^{t-k} p\left(c_{t-k+1}^{t} | c_{t-k-j+1}^{t-k}\right) \cdot \alpha[t-k][j] \qquad (4)$$

ただし $\alpha[0][0] = 1$ である．ここで式（4）の右辺は，対象とする部分文字列の前の可能なすべての分割により生成される部分文字列から，この部分文字列への遷移について周辺化していることを示す．

3.1.2 Backword sampling

Forward filtering で求めた $\alpha[t][k]$ から，最適な k を取り出すのが Backward sampling である．Backward sampling は文末から後ろ向きに k を取り出していく．そのため $\alpha[T][k]$（ここで T は文字列 s の長さを示す）から開始する．**図 3** に示すように，k は次式に比例して取り出し，文頭まで繰り返す．

$$k \propto p\left(w_i | c_{t-k+1}^{t}, \Theta\right) \cdot \alpha[t][k] \qquad (5)$$

ここで，w_i はすでに抽出した単語，Θ は言語モデルを示す．この手続を終えると $w_i, w_{i-1}, \ldots, w_1$ という単語分割が得られる．本研究では言語モデルとして NPYLM を用いている．

3.2 Nested Pitman-Yor Language Model

NPYLM は階層 Pitman-Yor 言語モデル（Hierarchal Pitman-Yor Langurage Model: HPYLM）[3] を拡張し，

図 4 検索ページ

図 5 検索結果一覧表示

文字 n-gram と単語 n-gram を組み合わせて表現した n-gram 言語モデルである．HPYLM は階層 Pitman-Yor 過程にもとづく n-gram 言語モデルである．単語を w，w の履歴を h としたとき，HPYLM における n-gram の条件付き確率 $p(w|h)$ は次式で求めることができる．

$$p(w|h) = \frac{c(w|h) - d \cdot t_{hw}}{\theta + c(h)} + \frac{\theta + d \cdot t_{h\cdot}}{\theta + c(h)} \cdot p(w|h') \quad (6)$$

ここで，$c(\cdot)$ は出現頻度，h' は h の 1 次元減らした履歴，d および θ は Pitman-Yor 過程におけるパラメータ，t_{hw} は HPYLM におけるパラメータ，$t_{h\cdot} = \sum_w t_{hw}$ を示す．NPYLM では単語 n-gram および文字 n-gram のそれぞれについての HPYLM を保持し，単語 n-gram における 0-gram の確率分布として文字 n-gram を用いる．

4. システム概要

本節では，本研究においてプロトタイピングした日本史史料に対するテキスト検索システムの概要を述べる．

構築したテキスト検索システムは非常に単純である．機能として，

(1) クエリ-史料テキスト間の類似度に検索結果のソート，
(2) 関連する史料の提示，
(3) 提示した史料を時系列に表示する可視化

を有する．(1) について，クエリ-史料テキスト間の類似度は式 (2) により求める．(2) においても式 (2) により求めた史料間の類似度を用いて提示する．(3) は，類似する史料をタイムライン上に配置する機能である．これにより，類似する史料と時系列の関係を分析できると期待する．

本システムでは日本の南北朝期（元弘 3 年～明徳 3 年 (1333～1392)）の史料を対象とした．テキストは SHIPS における大日本史料総合データベース，古記録フルテキストデータベース，古文書フルテキストデータベース，平安遺文フルテキストデータベース，および鎌倉遺文フルテキストデータベースから抽出し，7,007 史料，文字の異なり数が 4,067，延べ文字数が 1,204,594 であった．

図 4 は検索ページである．現在はテキスト内の文字列の検索しか行えない．そのため，かなりシンプルなページとなった．

図 5 は検索結果一覧画面である．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．ここでは，検索結果を式 (2) にもとづきソートしている．

図 5 における検索結果を選択すると，**図 6** に示すように，選択した史料の名称や本文とともに関連する史料を提示する．ここでの関連史料のソートも式 (2) により算出した類似度にもとづいている．

図 6 から “to timeline” をクリックすることで**図 7** 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[*6] 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

図 7 で関連する史料を確認することで，史料の類似性の時系列的関連を簡単に確認することができる．この仕組自体はあまり珍しいわけではないが，史料の類似性をテキストから抽出した単語をベースに行なっているケースは殆ど

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

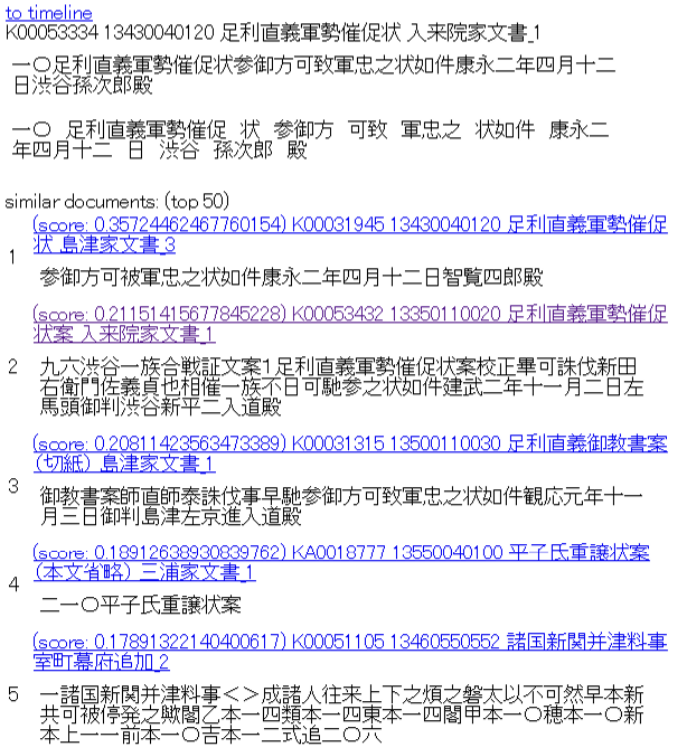

図 6 関連史料表示

図 7 タイムライン表示

無い．

単語分割の他の例を**図 8** に示す．図 1 においても共通して，日付部分の単語分割は，“日” や元号の直後がうまく分割できない場合もあるが，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” のような常套句も問題なく分割できていた．ただし，人名の直後に “申” が出現した場合，人名と連結してしまい，うまく分割できていない．また，“地頭” や “公文” の直後に “職” が出現すると，“職” は直後の文字と連結して

13750880550 赤堀文書 大日本史料6_46

text: 譲與所領事右美作国塩湯郷壱分地頭職并公文職内半分者兄義季観応二年二月廿一日戴御判今半分事康季同日戴御判者也彼御判同御施行遵行軍忠之御判等相副之甥為帯刀左衛門尉季治於猶子譲與者也聊一族親類中不可有違乱妨儀也仍為後日譲状如件永和元年八月十日藤原康季

Segment: 譲與 | 所領 | 事右 | 美作国 | 塩湯郷壱分 | 地頭 | 職并 | 公文 | 職内 | 半分者 | 兄義季 | 観応二 | 年二月 | 廿一日 | 戴御判 | 今半分事 | 康季同日 | 戴御 | 判者也 | 彼 | 御判 | 同 | 御施行 | 遵行 | 軍忠之 | 御判 | 等相副之 | 甥為 | 帯刀 | 左衛門 | 尉季 | 治於 | 猶子 | 譲與 | 者也 | 聊一 | 族親 | 類中 | 不可有 | 違乱 | 妨儀也 | 仍為後日 | 譲状 | 如件 | 永和元 | 年八月 | 十日 | 藤原 | 康季

13900080170 大内義弘挙状案 吉川家文書_2

text: 吉河讃岐入道玄龍申安芸国山県郡内志路原庄事依軍忠預置之候御下文事申御沙汰候者可然候恐惶謹言明徳元年八月十七日左京権大夫義弘在判進上御奉行所

Segment: 吉河 | 讃岐 | 入道 | 玄龍申 | 安芸国 | 山県郡内 | 志路原庄事 | 依 | 軍忠 | 預置 | 之候 | 御下文 | 事申 | 御沙汰 | 候者 | 可然候 | 恐惶 | 謹言 | 明徳元 | 年八月 | 十七日 | 左京 | 権大夫 | 義弘 | 在判 | 進上御 | 奉行所

図 8 単語分割の例

しまう傾向にあった．分割について学習する機能を追加することで，向上すると考えている．例えば，網羅はしていないが，ある程度整えられている人名辞典や地名辞典を学習に用いる，ユーザからのフィードバックを反映するなどの仕組みである．

また，古文書と古記録の間では類似度が低下してしまう傾向にあることがわかった．これは古文書と古記録の形式の差異によるところが大きいためだと考えている．そこで，人名や地名などのように史料の内容を特徴付ける上で重要だと考えられる用語の重みをより大きくすることで改善すると考えている．

また，人名において，同じ人物であるのにもかかわらず，家名，本姓，実名が記述されている場合と，通称のみが記述されている場合がある．また，本姓，実名は記述があるが，家名がない場合もある．このような場合，意味的にはある個人がテキスト内に 1 度出現したことになるが，家名，本姓，実名の記述がある場合，それぞれ 3 つの別の単語がそれぞれ 1 回ずつ出現することになる．そのため本姓，実名の記述がある場合とは出現する単語の頻度が異なってしまう．また，通称のみの場合は，もはや別の人物として扱われることになる．そのため，潜在的な解析を行う必要があると考えている．それを実現するためには，トピックのような潜在的な意味解析やシソーラスなどを用いた人物同定などの処理が必要となる．これは地名についても同様である．

6. おわりに

本稿では，南北朝期の日本史史料テキストに対する検索・分析を支援するため，ベクトル空間モデルを用いた類似史料の提示，タイムラインを用いた可視化の機能を有するテキスト検索システムをプロトタイピングし，その概要を述べた．また，ベクトル空間モデルを利用する上で欠くことのできない用語抽出のためノンパラメトリックベイズ手法である NYPLM を用いた．

今後，人名辞典や地名辞典などを用いた半教師あり学習（semi-supervised learning）やユーザからのフィードバックを取り入れる手法により単語分割の精度を向上させていく予定である．また，トピックのような潜在的情報を解析できる仕組みを導入し，史料の類似性についてもその精度を向上させていく予定である．

謝辞 研究の一部は，日本学術振興会科学研究費基盤研究（A）（23240031）の助成を受けたものである．

参考文献

[1] Gilks, W. R., Richardson, S. and Spiegelhalter, D. J.: *Markov chain Monte Carlo in practice*, Chapman & Hall, 1st ed edition (1996).
[2] Salton, G. and Buckley, C.: Term-weighting approaches in automatic text retrieval, *Information Processing and Management: an International Journal*, Vol. 24, No. 5, pp. 513–523 (1988).
[3] Teh, Y. W.: A Hierarchical Bayesian Language Model based on Pitman-Yor Processes, *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pp. 985–992 (online), available from ⟨http://www.aclweb.org/anthology/P/P06/P06-1124⟩ (2006).
[4] 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).
[5] 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（オンライン），入手先 ⟨http://ci.nii.ac.jp/naid/110008003480/⟩ (2010).