Vol.2010-IFAT-100 No.2
2010/11/4

古典史料における人名・地名を用いたテキストマイニング

大崎 隆比古　　井坪 将
立命館大学 理工学研究科

木村 文則　　手塚 太郎　　前田 亮
立命館大学 情報理工学部
〒525-8577 滋賀県草津市野路東 1-1-1
E-mail:cm001061@ed.ritsumei.ac.jp

概要

本論文では，古文書や古記録などの古典史料の本文に対してテキストマイニングを行う手法を提案し，古文書・古記録研究を行う上での活用手法とその有用性について述べる．本研究では，特に人名と地名の関係に着目し，その情報をディジタル化された古典史料のテキストデータから取り出し，ある人名と様々な地名の共起情報を，その人物固有のベクトルとみなし，そのベクトルの類似度を測ることで人物間の関係を可視化するシステムの構築を目指している．

キーワード

古典史料，テキストマイニング，可視化

Text Mining Using Place and Personal Names Extracted from Historical Documents

Takahiko OSAKI　Sho ITSUBO
Graduate School of Science and Engineering, Ritsumeikan University

Fuminori KIMURA　Taro TEZUKA　Akira MAEDA
College of Information Science and Engineering, Ritsumeikan University
1-1-1 Nojihigashi, Kusatsu, Shiga, 525-8577, Japan
E-mail:cm001061@ed.ritsumei.ac.jp

Abstract

This paper describes one of the ways of getting the most out of using historical materials and technical utility to use text mining. Our proposed method uses personal and place names

extracted from historical materials. We use the co-occurrence information of a personal name and various place names, and create a vector for the person based on that information. This paper aims at building a system that visualizes relations between persons by calculating similarities between these vectors.

Keyword
Historical documents，Text mining，Visualization

1. はじめに

ディジタル技術の発展は，いわゆる理系の学問分野のみに限らず，あらゆる分野で使われている．近年では，文系の学問分野においても，そのデータ管理等にディジタル技術を用いて行っている例は少なくない．特に歴史的な書物や遺跡などは，経年劣化等により破損が進むなど保存方法が問われている．史料の長期保存を目的として，近年ではディジタル技術を用いた保存もひとつの方法として大いに進められている．例として，現在，国立美術館が，平成 17 年度末までに収蔵した所蔵作品の総合目録を作成し公開している独立行政法人国立美術館所蔵作品総合目録検索システム[1]や，日本芸術文化振興会の文化デジタルライブラリー[2]などが挙げられる．

このように，近年では，史料のディジタル化も大きく広まりつつあるといえる．しかし現在では，古典史料の保存がその主たる目的となっているのが現状である．上記システムにおいても，作者やページ数など史料に関連するデータを扱う簡素なものがほとんどであり，古典史料の内容そのものに踏み込むようなシステムは，あまり開発されていないのが現状である．実際，文学研究において，保存形態の一つとしてこそディジタル技術は用いられているが，未だほとんどの研究が手作業で進められている．

そこで本研究では，ディジタル化された古典史料テキストから，有用なデータを抽出する．ここでは特に地名と人物との間の共起関係に着目し，それらを集積し処理を行うことで，地名を通して人物間の関連性を得る手法を提案する．

2. 古記録『兵範記』[3]

『兵範記』とは，平安時代後期の貴族，平信範（たいらののぶのり，1112~87）が記した日記である．平信範の漢字から,『人車記』や『平洞記』などとも呼ばれる．天承二年（1132）から元歴元年（1184）まで記録されたことが伝わっているが，現在では天承二年から承安元年（1171）までの約 40 年間分である自筆浄書本 54 巻のみが現存している．また，現存している巻も，破損，汚損，劣化等により完全な形では残っておらず，ばらばらになり，どの巻のどこにあたるのかわからない断簡と呼ばれる物も存在する．

平信範は朝廷に長期間勤め，鳥羽・後白河院の院司や，摂関家の藤原忠通，藤原基実らに

Vol.2010-IFAT-100 No.2
2010/11/4

家司として使えるなど当時の政治の中枢にいた人物である．また，『兵範記』は日記であるが，日頃身近に起こったことをただ記述する現代の日記とは異なり，当時の日記は公事や政治の動きなどを日記として記述するものであった．そのため，『兵範記』には，政策決定に至る推移や行政文書の写し，要人の見解，朝廷・院・摂関家に関する儀式次第など，当時の朝廷や政治の様子が詳細に描かれている．このため，歴史資料として価値が高い史料となっている．京都大学電子図書館[4]で公開されている『兵範記』の自筆・古写本の第一巻『長承元年秋冬』の画像データを図 1 に示す．

本研究においては，『兵範記』刊本を基に，立命館大学文学部杉橋研究室で電子テキスト化し，校訂されたものを用いる．

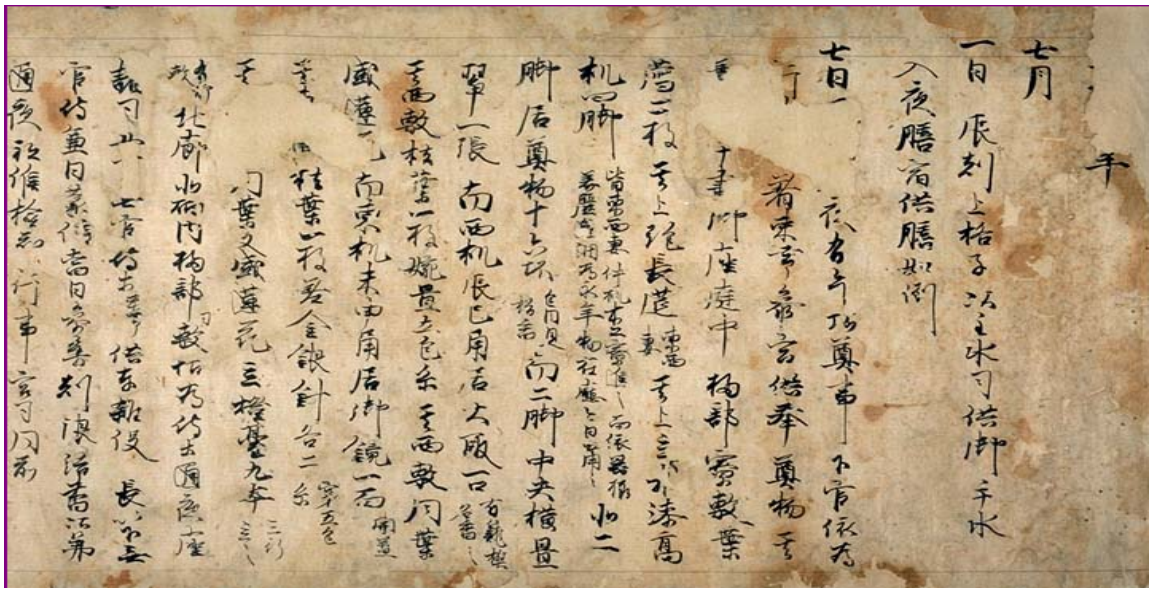

図 1:『兵範記』の自筆・古写本の第一巻『長承元年秋冬』の画像データ

3. 関連研究

文学研究において保存用途以外に，ディジタルテキストとして保存されているデータの活用を目的としたシステムは，ほとんど公開されていない．我々はこれまでに『兵範記』のテキストデータに人名及び単語情報の付加や，『日付指定検索』などの検索機能を含めた検索システムの開発などを行っている[5].

また，文学研究においては，『兵範記』の地名，人名，その他天気などの情報を取り出し，移動経路から当時の様子を探る研究が行われている[6]．吉田の研究では，『兵範記』およびその他いくつかの古記録から，特定の人物が移動した経路を調べ，それを集計し，官位や地位といったもの違いや，時間の経過に伴う生活の変化を移動経路から読み取る試みを行っている．これらの結果から，当時の生活や習慣を読み取る上で，地名等を集計し適切な処理を加え観察を行うことは，非常に有意義な模索であることを示している．しか

しながら，これらの作業は全て手作業で行われており，新しい要素を追加する，あるいは新たな古記録の解析を行うなどした場合など，さらに多くの時間を要する．

このような状況を考慮に入れ，本研究では『兵範記』より抽出した人名及び地名データより，それらのデータに処理を施すことによって人物間の関係性を取得し，それを可視化するシステムの構築を目指す．

4. 古典史料からの情報抽出および人物間の関係性の可視化

4.1 情報抽出

本研究のシステムでは，人物間の関係性を獲得するために，その人物と関係のある地名を取得する．『兵範記』において，文章は漢文体で書かれている．そのため，一般に文章を解析する際に用いられる形態素解析は，『兵範記』には用いることができない．そこで本研究では，人名と地名にしぼって情報を抽出する．地名データは『平安京提要』[7]，『京都市の地名』[8]やその他の文献から立命館大学文学研究科の谷昇氏によって作成された『京都地名索引』を用いる．また人名に関しては『兵範記人名索引』[9]を基に，出現箇所を求める．

人名と地名の関連を表す情報としては共起頻度を用いる．その際の共起の範囲は，一段落内での共起とする．漢文体の文章には，基本的には句読点も段落も含まれていないのであるが，ディジタル化された『兵範記』の文章には，文の切れ目に句点が，話の切れ目に段落がそれぞれ付加されている．そこで本研究では，この情報を用いることで，日ごとの共起より正確な共起情報を得ることができる段落を共起の範囲とした．

4.2 人物間の関係性の獲得と可視化

前節で述べた情報抽出の処理で得た地名との共起情報をもとに，人物間の関係性を獲得する．

関係性を獲得するために，情報抽出で得た各地名をそれぞれ一つの次元とし，地名数の次元を持つベクトルを作成し，そのベクトルを基に人物間のコサイン類似度をとることで，人物間の関係性を得る．

また得られた類似度を基に，K-means 法を用いて人物をクラスタリングすることで，人物のグルーピングを行う． K-means 法では，初期値の割り振り方によってクラスタリング結果が異なるが，初期値を確率的にまんべんなく与えることでクラスタリングが比較的短時間で最適解に近似しやすくなるという報告がある[10]．そこで，本研究では確率的に初期値を与えることでクラスタリング結果を得る手法を用いる．今回のシステムで用いている K-means 法の手順を以下に示す．

(1) 初期値 x をランダムで選択する(x はある人物の地名との共起から得たベクトル)

(2) $p(x) = \frac{D(x')^2}{\sum D(x)^2}$ を最大にする x' を新しいクラスタ中心として追加する($D(x)$ はデータ x から一番近いクラスタまでの距離．ここでは類似度を 1 から引いた数を距離として扱う)

(3) (1),(2)の過程を指定したクラスタ数になるまで繰り返す

(4) すべてのデータを一番近いクラスタ中心を持つクラスタに割り振る

(5) 各クラスタ中心を割り振られたデータから算出する

(6) データのクラスタ間の移動がなくなるまで(4)，(5)を繰り返す

(7) (1)から(6)までの作業を指定回数まで繰り返し,各データから一番近いクラスタ中心までの距離の総和が最少となったものを，クラスタリング結果として取得する

ここまでで得た，人物間の類似度とクラスタリング結果を用いて可視化を行う．可視化は JUNG[1]を用いた可視化プログラムによってグラフ化される．JUNG とは，Java で記述されたオープンソースライブラリであり，グラフで表現できる情報の可視化や解析を行うフレームワークである．

5. 実験

5.1 実験の内容

4 章で示した手法で獲得した人物間の関係を可視化したグラフの作成を行った．

今実験では，対象とする期間を，保元の乱がおこった 1156 年の 7 月の初めから 7 月の末までの間に絞って実験を行った．保元の乱とは，地位をめぐる確執から後白河天皇と兄の崇徳上皇が対立し，双方の武力衝突に至った政変のことである．可視化を行う人物は，上皇に従ったものと，天皇に従ったもののなかから任意で選んだ 78 人を対象としている．

上記の期間において，78 人の人物で実験を行い，実際に『兵範記』から地名データを取り出すことができた 31 人の関係について可視化を行っている．クラスタリングでは，クラスタ数は 3，初期値を求めるループ数は 20 として実験を行った．

図 2 に 31 人について関係性の可視化を行ったグラフを示す．ここではグラフのラベルである名前の横に付いている数字が 1 のものは上皇派で色が緑色，2 のものが天皇派で色が赤色で表示されている．また図 3 には，図 2 同様の 31 人の関係性の可視化を行ったものに，クラスタリング結果の情報を，図 2 同様にラベルの名前の後に数字として付加したものを示す．それぞれ 1 つ目のクラスタは緑，2 つ目のクラスタは赤，3 つ目のクラスタは青で表示される．表 1 には，上皇派か天皇派か，どのクラスタに分類されたかをまとめたものを示す．また，グラフの破線は人物間の類似度が 0.4 から 0.7 まで，0.7 以上の場合は実線で

[1] http://jung.sourceforge.net/

描かれている.

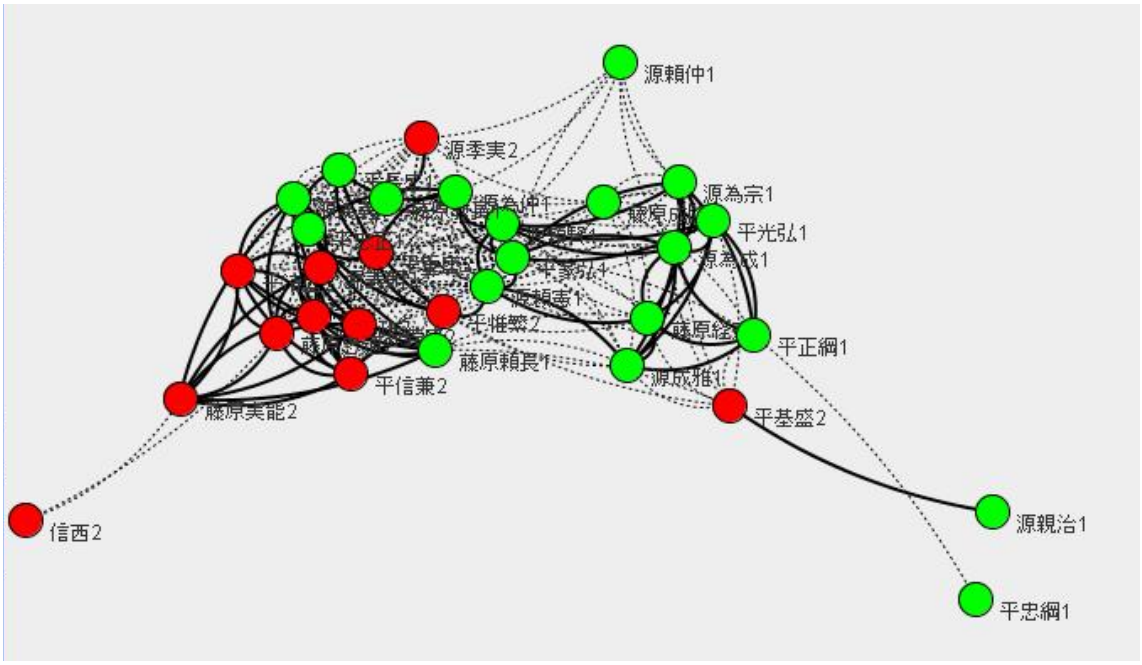

図 2:保元の乱，人物間の類似度から可視化したグラフ

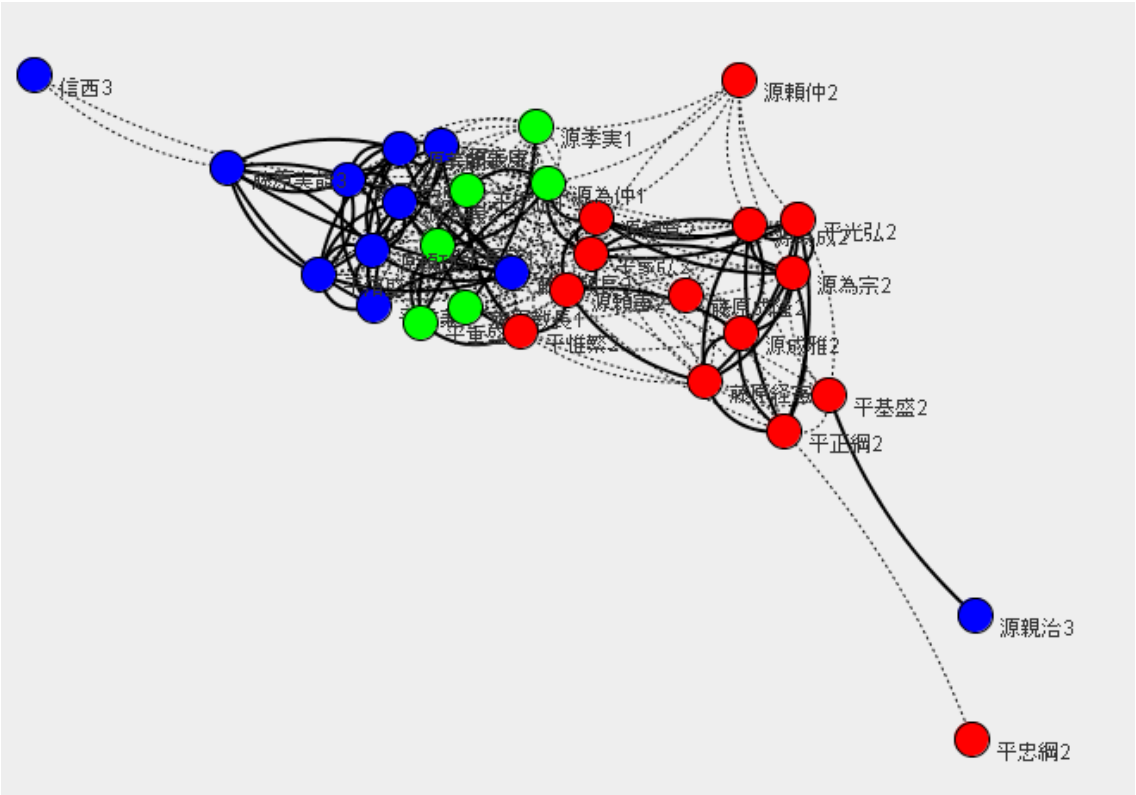

図 3:保元の乱，人物間の類似度から関係を可視化した
グラフにクラスタリング結果を付加したグラフ

表 1:保元の乱，クラスタリング結果及び上皇派，天皇派の表

	クラスタ 1	クラスタ 2	クラスタ 3
上皇派	平長盛，藤原教長，源為仲，平忠正，	源成雅，平忠綱，源頼賢，源頼仲，源為成，源頼憲，藤原経憲，平正綱，源為宗，平家弘，平光弘，藤原成隆	藤原頼長，源親治，
天皇派	源季実，平重盛，平惟繁	平基盛	信西，源義朝，源頼政，源為義，藤原実能，平清盛，源義康，藤原忠通，平信兼

5.2 実験結果の考察

本実験では，人物と共起した地名を，その人物特有の特徴として人物間の類似度を求めクラスタリングを行った．その結果，非常に特徴を持ったグラフの作成に成功している．図 2 では大まかに分けると右側を上皇派，左側を天皇派とで分けることができる．実際にクラスタ数を 3 でクラスタリングした図 3 をみても，左右で二つに分かれ，中心の上皇派と天皇派が少し入り乱れている部分とで 3 つに分類されている．

クラスタリング結果をみると，クラスタ 1 以外ではおおむね上皇派と天皇派で別れているが，クラスタ 3 において，上皇派の中心人物である藤原頼長が天皇派と一緒にクラスタリングされている．

6. おわりに

今研究では，『兵範記』から得られた地名と人名の共起情報を基に，人物間の関係の可視化と人物のクラスタリングを行った．その結果，上皇派，天皇派とその中間にあたるグループにクラスタリングを行うことができた．クラスタリングを行うことで，それぞれの人物がどの立ち位置にあったのかなどが非常に分かりやすくなり，またそれを可視化することによって，非常に直観的でわかりやすいものになったといえる．

今後の課題としては，クラスタリングの精度の向上が挙げられる．今回のように人物数が少ない場合ではクラスタ数は少なくて済み，またクラスタ数の決定も直観的なものでも十分であるが，人数が多くなるとクラスタ数を直観的に設定するのは非常に難しくなる．K-means 法のクラスタをさらに分割･併合するアルゴリズム[11]等を用いることによって，分類誤りの減少やクラスタの数も決定しやすくなると思われる．

また，今回の実験で人物と地名の共起が，端的にその人物の特徴を十分に表わし得ることがわかったので，今後はそれを時間軸に沿って解析するなど，日記の特性を生かした手法の模索も課題の一つである．

謝辞

本研究の一部は文部科学省グローバル COE プログラム「日本文化デジタル・ヒューマニティーズ拠点」，文部科学省私立大学戦略的研究基盤形成支援事業「芸術・文化分野の資料デジタル化と活用を軸とした研究資源共有化研究」，文部科学省科学研究費補助金若手研究(B)「言語・時代・文化横断型の情報アクセスに関する研究」(研究代表者:前田亮, 課題番号:21700271) の支援を受けている.

参考文献

[1] 独立行政法人国立美術館，独立行政法人国立美術館所蔵作品総合目録検索システム，http://search.artmuseums.go.jp/

[2] 日本芸術文化振興会，文化ディジタルライブラリー，http://www2.ntj.jac.go.jp/dglib/

[3] 兵範記:京都大学文学部国史研究室、思文閣出版、1988.

[4] 京都大学付属図書館，京都大学電子図書館，http://edb.kulib.kyoto-u.ac.jp/minds.html

[5] 木村 文則， 小牟礼 雅之， 前田 亮， 佐古 愛己， 杉橋 隆夫: 古典史料データベース検索システムの提案，情報処理学会研究報告，2008-CH-78，pp.45-52，2008.

[6] 吉田真澄:貴族の移動経路から見た平安京，第 29 回立命館大学グローバル COE GCOE セミナー，2008/7/28，http://www.arc.ritsumei.ac.jp/lib/GCOE/seminar/asx2/20080729yoshida.asx

[7] 平安京提要: 古代学協会・古代学研究所，角川書店，1994.

[8] 日本歴史地名大系 27 京都市の地名: 林屋辰三郎，村井康彦，森谷尅久，平凡社，1979.

[9] 兵範記人名索引:兵範記輪読会，思文閣出版，2007.

[10] 坂井美帆，山田誠二，小野田崇:独立成分分析による k-means 法の初期値設定手法の提案，2010 年度人工知能学会全国大会（第 24 回）, 2010.

[11] 倉橋 和子，森井 藤樹:分割・併合機能を有する K-Means アルゴリズムによるクラスタリング，電子情報通信学会技術研究報告. PRMU, パターン認識・メディア理解 106(470), 67-71, 2007.