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　２９　年　　６　月　２３　日現在

機関番号：　２５４０５
研究種目：　若手研究(B)
研究期間：　2014～2016
課題番号：　２６７３０１６９

研究課題名（和文）古典資料に対するテキストマイニングおよびその分析結果の可視化に関する研究

研究課題名（英文）Research of text mining for historical documents and visualization of analysis result

研究代表者

木村　文則（Kimura, Fuminori）

尾道市立大学・経済情報学部・講師

研究者番号：７０５１６６９０

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：　本研究ではコンピュータによる古典史料を対象にした知識獲得の手法を確立する手法を提案し，一定の成果が得られた．
　重要な古文の単語の抽出器の作成は，SVMを用いた手法を提案し，自動的に抽出できる仕組みを構築した．古典史料からの知識獲得および可視化方法については，人物関係図を可視化するシステムを構築し，動的に操作できるようにすることにより分析の支援ができるようにした．また，地名を媒介とした人物関係を分析し，クラスタリングを行うことにより，同時代でない人物同士の類似性も測ることができるようになった．これにより，コンピュータを用いて大量の古典史料から知見を得るための支援システムを実現することができた．

研究成果の概要（英文）：I have to realize the methods in order to achieve research of information extraction and visualization from ancient documents.
I proposed the method using support vector machine in order to extract important terms from historical documents, and constructed its system. I also constructed the system to visualize relationships between persons. This system can be used dynamically in order to support analyzing extracted personal relationships. I define that having a strong personal relationship means that the persons have a similar relationship with the places because we believe that one's personal status is strongly related with the places they frequent. This method can estimate the relationships among persons who do not necessarily have direct relationships.

研究分野：　情報検索

キーワード：　テキストマイニング

１．研究開始当初の背景

人文科学の領域においては，古典史料を調査することにより知見を得ることは，非常に重要な研究課題である．これまでは，人文科学の研究者が多大な労力をかけ，人手により丹念に古典史料を調査することにより,研究成果を積み重ねてきた.しかし，人手による調査には，量的な限界が存在するため，調査が行き届いていない古典史料が数多く残っている．また，大量の古典史料を人手により網羅的に解析することも非常に困難である．

近年，コンピュータを様々な分野に活用することが注目されている．人文科学の分野においても同様である．コンピュータは大量のデータに対して処理を行うことに優れているため，複数の古典史料に対して網羅的に解析を行うことができる可能性を秘めている．

コンピュータに古典史料を解析させるためには，古典史料が電子テキスト化されている必要がある．1990 年代半ば以降の Web の世界的な発展により,情報の電子テキスト化が活発に行われるようになり，電子化された古典史料も徐々に増加しつつある．このように，大量の古典史料に対する解析をコンピュータにより行うための環境が整い始めたことから，人文科学分野でもコンピュータを用いて古典史料を解析することが行われ始めている．

しかし残念ながら，古典史料に対するコンピュータによる解析手法はまだ研究されていない．そこで本研究では，大量の古典史料に対する解析をコンピュータにより行うための手法の提案を行う．大量の文書に対してコンピュータによる解析を行う技術として,「テキストマイニング」がある．これまでテキストマイニングは，現代語で記述された文書に対して適用されてきた．テキストマイニングの技術を古典史料にも適用できるようにすることにより，コンピュータを用いて大量の古典史料から知見を得るのが本研究の目的である．また，得られた知見を可視化することにより，古典史料を研究するための支援を行う．

２．研究の目的

本研究では，古典史料に対する情報抽出およびその可視化手法の提案を行う．電子テキスト化された古典史料から情報の抽出を行い，それらの分析をし，その結果の可視化を行う．特に，時間の経過に応じて変化している関係・あるいは変化しない関係が直感的に視認できる可視化方法の提案を行う．

古典資料に対してテキストマイニングを行うためには，まず，古典史料から「人物名」,「地名」,「事柄」などの情報を抽出する必要がある．抽出した情報に対しテキストマイニングを行うことにより分析を行い，人物同士の関連などを導き出す．記述された日が明記された古典史料を対象とすることにより，人物同士の関連を時系列に追跡することが可能であり，その解析手法の提案を行う．こうして得られた知見を提示するために適した可視化方法，特に時系列変化に注目した可視化方法を検討し，そのシステム構築を行う．

３．研究の方法

本研究では，古典史料に対する情報抽出およびその可視化手法の提案を行う．本研究では，平安時代から鎌倉時代にかけて成立した古典史料を対象とする．日本語の古文で書かれたものだけでなく,『兵範記』,『吾妻鏡』などの漢文体で書かれたものも対象とする．

本研究を実現するために，以下の流れで研究を進める．

1. 古文に対してテキストマイニングを行うために必要となる重要な古文の単語（人名,地名など）の抽出器の作成を行う．

2. 1.で作成した言語資源を利用し，古典史料からの知識獲得を行う．

3. 2.で得られた知見を可視化し，提示するシステムの構築を行う．特に，時系列による変化が認識しやすい可視化方法の提案を行う．

４．研究成果

(1) 古典史料からの知識獲得および可視化

古典史料の本文データから，人物と地名の共起情報を用いることで人物関係の可視化を行う手法を提案した．また，可視化の際に生成した人物関係ネットワークを分析することにより，関係性の大きい人物をクラスタリングする手法を提案した．さらに，生成された各クラスタの特徴を，そのクラスタに属する人物群との関連の大きい地名情報を取得することにより表現することを提案した．

解析対象として古典史料のディジタルデータの一つである平安時代末期の日記である『兵範記』を用いる．人物関係の取得には,『兵範記人名索引』，地名の取得には『京都地名索引』という，人手により作成された関連史料を用いることによって,『兵範記』本文中から取得した人物と地名の共起頻度を用いる．取り出した人物と地名との共起頻度を基に，各人物の特徴をベクトルとして表現し，そのベクトルから人物間の類似度の評価を行い，人物関係図を作成する．これにより視覚的に人物関係を捉えることができる仕組みの提案を行った．図 1 は,『兵範記』に記述されている平家の人物の関係性を可視化した結果である．

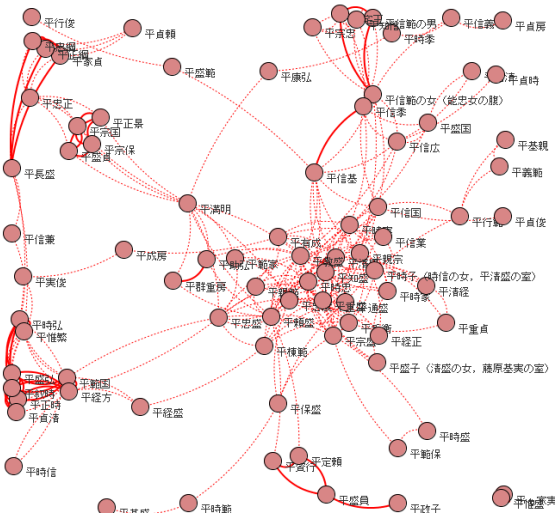

図 1. 『兵範記』に記述されている平家の人物の関係性

また，得られた人物関係図を基に，Girvan-Newman クラスタリングアルゴリズムを用いて人物をクラスタリングすることで，人物のクラスタリングを行った．Girvan-Newman クラスタリングアルゴリズムは，ソーシャルグラフに対するハードクラスタリング手法の階層的クラスタリングである． 同じクラスタ内の他のノードとの関連の弱いエッジから順に切断してゆき，ある任意の数のエッジを切断した段階で切断を終了する．その結果，関係性の近い人物同士が同一のクラスタとしてクラスタリングされる．図 2 は，図 1 の人物関係図を上記手法でクラスタリングした結果である．図 2 では，42 本のエッジを切断した．

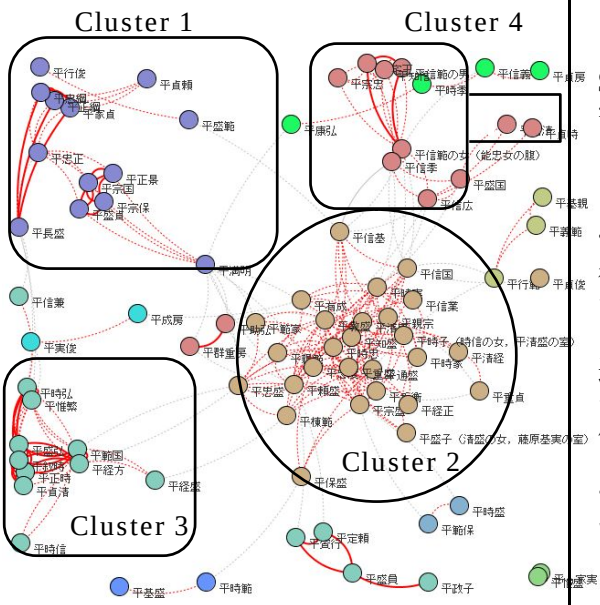

図 2. 『兵範記』に記述されている平家の人物のクラスタリング結果

図 2 では，主に 4 つの主要なクラスタが生成されている．この主要な 4 つのクラスタに属している人物が適切にクラスタリングされたかどうかを，歴史的事実を基に判定した結果を表 1 に示している．判定が困難な人物を除外した場合のクラスタリング精度は，平均で約 89%となっており，本手法が有効であることが示されている．

表 1. 『兵範記』に記述されている平家の人物のクラスタリングの精度

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	平均
全人物	0.3636	0.9231	0.5455	0.6000	0.6897
判定可能な人物	0.6667	1.0000	0.8571	0.7500	0.8889

本手法の特徴は，人物同士の関連を直接的な関係ではなく，地名を介した間接的な関係性を基に抽出していることである．『兵範記』の書かれた平安時代では，地名と人物のつながりは現代よりも強く，地名は当時の人物の官位，家柄，役職などの特徴を表す重要な要素であるといえる．本手法は，この点を考慮した分析に適している．また，地名を介することにより，同時代でない人物同士の類似性も測ることができることも特徴である．

(2) 古典史料からの重要な古文の単語（人名，地名など）の抽出器の作成

解析対象となる古典史料を拡大するためには，人手を介する箇所を減らし，自動化することが望ましい．そこで，ある程度の学習データを用いて機械学習を行い，重要な語を自動で取得する手法を提案した．本手法では Support Vecter Machine(SVM)を用い，各文字の周囲の文字とその出現パターンを学習し，それを利用して重要な語を抽出する手法を提案した．

本研究においては，12 世紀ごろの歴史史料である『東大寺要録』および，江戸時代に刊行された歌舞伎役者の評判について書かれた『役者評判記』を対象とした．

『東大寺要録』は，もともとは漢文体で記述されているが，『東大寺要録』の研究グループが書き下し文への翻刻および一部注釈作業を行っている．

本手法では上記の注釈を学習データとして用い，書き下し文を対象に，新たに注釈とすべき語を抽出した．図 3 は，その概要である．

また，SVM の素性データとして，文字の種類（漢字，ひらがな）および単語の境界情報を用いた．単語の境界情報は，文字 n グラムの出現確率を基に自動的に単語境界を推定する手法を用いている．この手法により，注釈の取得の適合率 0.8411，再現率 0.7171，F

値 0.7742（部分一致も正解とした）という結果が得られた.

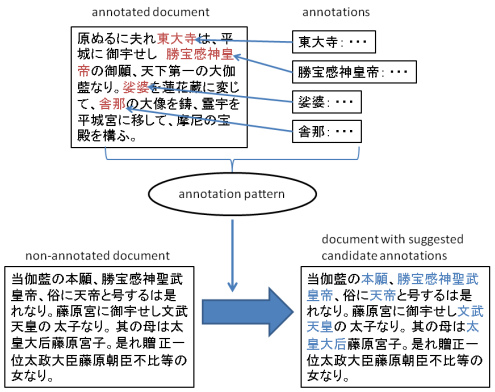

図 3. 『東大寺要録』からの注釈の抽出の概要

『役者評判記』も，基本的には『東大寺要録』の場合と同様の手法で重要語の取得を行っている．ただし，単語境界情報は，中古和文 UniDic を用いて取得している．本手法により，人物情報にに関しては適合率 0.9289，再現率 0.9100，F 値 0.9193 という結果が得られた.

本研究ではコンピュータによる古典史料を対象にした知識獲得の手法を確立する手法を提案し，一定の成果が得られた.

重要な古文の単語（人名，地名など）の抽出器の作成は，SVM を用いた手法を提案し，自動的に抽出できる仕組みを構築した．人名の抽出などにおいてはよい精度が得られた.

古典史料からの知識獲得および可視化方法については，人物関係図を可視化するシステムを構築し，動的に操作できるようにすることにより分析の支援ができるようにした．また，地名を媒介とした人物関係を分析し，クラスタリングを行うことにより，同時代でない人物同士の類似性も測ることができるようになった．これにより，コンピュータを用いて大量の古典史料から知見を得るための支援システムを実現することができた.

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 1 件）

① 著者名：Takafumi Sato、Makoto Goto、Fuminori Kimura、Akira Maeda、論文表題：Developing a Collaborative Annotation System for Historical Documents by Multiple Humanities Researchers、雑誌名：International Journal of Computer Theory and Engineering、査読：有、Vol. 8, No. 1, 発行年：2016 年、ページ：88-93

〔学会発表〕（計 6 件）

① 発表者名：Fuminori Kimura、Akira Maeda、発表表題：Method for Supporting Analysis of Personal Relationships through Place Names Extracted from Documents、学会名等：Digital Libraries 2014: ACM/IEEE Joint Conference on Digital Libraries (JCDL 2014) and International Conference on Theory and Practice of Digital Libraries (TPDL 2014)、発表年月日：2014 年 9 月 10 日、発表場所：ロンドン（イギリス）

② 発表者名：佐藤 貴文、後藤 真、木村 文則，前田 亮、発表表題：東大寺要録からの歴史知識情報の抽出 ―注釈情報の共有を目指して―、学会名等：人文科学とコンピュータシンポジウム、発表年月日：2014 年 12 月 13 日、発表場所：国立情報学研究所（東京都）

③ 発表者名：永井 規善、前田 亮、木村 文則、赤間 亮、発表表題：役者評判記からの人物表現抽出手法の提案、学会名等：人文科学とコンピュータシンポジウム、発表年月日：2014 年 12 月 13 日、発表場所：国立情報学研究所（東京都）

④ 発表者名：Takafumi Sato、Makoto Goto、Fuminori Kimura、Akira Maeda、発表表題：Developing a Collaborative Annotation System for Historical Documents by Multiple Humanities Researchers、学会名等：The 7th International Conference on Computer Science and Information Technology (ICCSIT2014)、発表年月日：2014 年 12 月 23 日、発表場所：バルセロナ（スペイン）

⑤ 発表者名：Takafumi Sato、Makoto Goto、Fuminori Kimura、Akira Maeda、発表表題：Extracting Key Phrases for Suggesting Annotation Candidates from Japanese Historical Document、学会名等：Digital Humanities 2015、発表年月日：2015 年 7 月 1 日、発表場所：シドニー（オーストラリア）

⑥ 発表者名：Noriyoshi Nagai、Fuminori Kimura、Akira Maeda、Ryo Akama、発表表題：Personal Name Extraction from Japanese Historical Documents Using Machine Learning、学会名等：The 5th International Conference on Culture and Computing (Culture and Computing 2015)、発表年月日：2015 年 10 月 18 日、発表場所：京都（日本）

６．研究組織

(1)研究代表者

木村 文則 （KIMURA FUMINORI）
尾道市立大学・経済情報学部・講師
研究者番号：70516690