古典史料からの知識獲得および情報の可視化

井坪 将　木村 文則　手塚 太郎　前田 亮
立命館大学 情報理工学部

近年，情報技術の進歩により，ディジタル化された古典史料の本文や現代語訳などの情報がインターネット上に公開され始めている．しかし，これらの情報は公開されているが，知識獲得のために用いられていない．そこで本論文では，古典史料の現代語訳の形態素解析を行い，古典史料に登場する人物を抽出し，人物同士の共起頻度を算出する．得られた共起頻度情報を散布図に表し，人物間の関係を表示する．さらに，散布図から知識獲得を行う．また，古典史料の原文と現代語訳に登場する人物間の共起頻度の類似度を算出する実験を行った．

Knowledge Acquisition and Visualization from Historical Documents

Sho Itsubo　Fuminori Kimura　Taro Tezuka　Akira Maeda

College of Information Science and Engineering,
Ritsumeikan University

Recent advances of information technology have made it possible to publish historical materials through the internet, including the original text in ancient language and in some cases with their translations in modern language. However, most of these contents are made primarily for making them open to the public, and are not intended for knowledge acquisition. In this paper, we propose a method to acquire knowledge about the relations of people which appear in a historical diary in ancient Japanese using its translation in modern Japanese. We first do morphological analysis of modern Japanese text, then extract personal names, and calculate co-occurrence frequencies between persons. We use scatter plots for visualizing the relations of people obtained from the co-occurrence information. We draw scatter plots for each year in order to acquire knowledge about changes in relations of people over time. Besides, we conducted an experiment of calculating similarities of co-occurrence frequencies of personal names between the original text and its translation in modern language.

1. まえがき

近年では，ディジタル技術の進歩から古典史料のディジタル化による保存が進んでいる．また，ディジタル化されたこれらの古典史料がインターネット上で公開され始めている．ディジタル化され始めた当初は，古典史料を保存することが主な目的であり，その史料のテキストや画像データ，著者や成立年代などの簡単な情報を公開することが多かった．最近では，徐々にではあるが，古典史料の本文や現代語訳をテキストデータ化し，公開する例が増えてきている．また，公開されたデータを用いて，データベースが作成されている[1]．しかし，これらのデータを解析し，それにより新たな知識を獲得することはほとんど行われていない．

本論文では，古典史料の現代語訳のテキストデータをテキストマイニングの技術により解析を行い，解析結果を可視化することによって，古典史料から知識を獲得する手法を提案する．解析結果を可視化する際に，理解しやすい提示方法も検討する．本論文では，鎌倉時代に成立した歴史書『吾妻鏡』の現代語訳[2]を用いて解析を行い，知識の獲得を目指す．

2. 関連研究

テキストデータを解析する技術にテキストマイニングがある．テキストマイニングの研究として，文書間や単語間の関連表示，単語間の相関ルールや時系列パターンの抽出などが行われており，情報抽出や相関ルールや可視化の技術が用いられている．これらの技術によって，テキスト中の必要な情報などを抽出し，抽出された情報の集計も行うことができる[3]．テキスト中から情報を抽出するためには，文字単位より単語単位の方が良い．自然言語の文法知識や辞書により，テキストを形態素に分割し，品詞情報を付与する形態素解析システムとして，Chasen や MeCab などがある．

本研究では，これらのディジタル技術を用いて，古典史料の現代語訳の解析を行い，情報抽出を目指す．

表 1：類似結果

人物名	源頼朝	北条時政	上総広常	大庭景親	土肥実平	以仁王
現代語訳の共起頻度	0	17	17	15	14	11
原文の共起頻度	0	12	10	9	12	7
標準化後の現代語訳の共起頻度	-0.794	4.340	4.340	3.736	3.434	2.528
標準化後の原文の共起頻度	-0.860	4.137	3.304	2.887	4.137	2.054
	平均	標準偏差				
現代語訳	2.629	3.310				
原文	2.065	2.401				
標準化前の相関係数	0.917					
標準化後の相関係数	0.909					

3. 原文と現代語訳の登場人物の出現傾向の類似比較実験

古典史料の原文から抽出したい情報を正確に取り出すことや解析を行うことは，難しいのが現状である．しかし，現代語に関しては，自然

言語処理の技術を用いることにより，解析などを行うことができる．そのため，古典史料を解析するには，古典史料の現代語訳が必要になり，古典史料の原文と現代語訳の情報が類似していることが重要である．本研究では，『吾妻鏡』の現代語訳を対象に実験を行うため，『吾妻鏡』の現代語訳と『吾妻鏡』の原文との類似を比較するために，『吾妻鏡』の現代語訳に登場する人物間の共起頻度と『吾妻鏡』の原文に登場する人物間の共起頻度の相関係数を算出した．相関係数を算出する実験手順は以下の通りである．

1．原文と現代語訳それぞれの共起頻度を算出
2．共起頻度の値を標準化する
3．標準化された共起頻度の相関係数を算出

原文と現代語訳に登場する人物間の共起頻度 X から原文と現代語訳の共起頻度の平均 μ と標準偏差 σ を算出し，これらの値を用いて共起頻度 X の標準化を行う．標準化変数 Z は，以下の式から算出される．

$$Z = \frac{X - \mu}{\sigma}$$

算出された原文と現代語訳の共起頻度 X の標準化変数 Z の相関係数を算出する．表 1 は，原文と現代語訳の共起頻度，標準化後の共起頻度，相関係数による原文と現代語訳の類似比較の結果の一部を示す．表 1 から標準化前の相関係数と標準化後の相関係数共に相関係数の値は非常に 1 に近い値のため，原文と現代語訳の登場人物は類似していると言える．さらに，現代語訳では原文には書かれていない人物を補って書かれているため，原文より正確な共起頻度を算出できると考えられる． この実験は，1180 年の人物間の共起頻度の中から数人に対して行い，表 1 は『吾妻鏡』の現代語訳と原文の源頼朝とその他の人物との共起頻度を用いて，実験を行ったものである．

4. 提案手法

本論文では，古典史料の現代語訳をテキストマイニングによって解析し，得られた結果を可視化する．その手法の一つとして，古典史料に登場する人物を対象に共起頻度から人物関係を表す散布図を年代ごとに作成することにより，人物間の関係の変化を提示することを提案する．

本手法では，人物間の共起頻度を基に人物間の関係の深さを算出し，その結果から散布図を作成することによって視覚的に表現する．

まず，現代語訳の形態素解析を行い，単語を抽出する．本論文では，固有名詞である『人名』を抽出し，人物の関係を示す．現代語訳の形態素解析を行うために，MeCab[4]というソフトウェアを用いて処理を行う．名前に複数の表記がある人物がいるため，抽出する名前を統一する必要がある．そのために，『吾妻鏡』の原文と人名索引が収録されている吾妻鏡データベース [5]から人名索引を用いて，人名ファイルを作成した．この人名ファイルは，各行に書かれている同一人物の複数の表記を各行の最初の名前に統一される．表 2 は，人名異表記ファイルを示す．

表 2：人名異表記ファイル

以仁王 以仁王 一院第二宮 高倉宮 一院第二皇子 茂仁王 宮 新王 新皇 三条宮 皇子 三条高倉宮
源為義 為義 廷尉 六条廷尉禅室 廷尉禅門 大夫尉 六条殿 故六条廷尉禅門
平佐古為重 為重 太郎
為貞 為貞 兵衛志
長尾為宗 為宗 新五 新五郎 為家
中原惟重 惟重 四郎 維重 是重
平維盛 維盛 惟盛 権亮少将 左近少将 小松少将 小松羽林 左少将 頭中将 権亮三位中将 故惟盛卿
維平 維平 惟平 中八 ちうはち
千葉胤信 胤信 常胤子息 四郎 常胤四男 胤通 大須賀四郎跡
千葉胤正 胤正 胤政 常胤子息 太郎 小太郎 常胤嫡男 新介

表 3：人物間共起頻度情報

	源頼朝	北条時政	平広常	大庭景親	土肥実平	以仁王	千葉常胤	佐々木定綱	源義朝	平清盛
源頼朝	0	17	17	15	14	11	10	10	9	9
北条時政	17	0	4	6	9	3	2	6	4	3
平広常	17	4	0	3	6	0	10	3	5	1
大庭景親	15	6	3	0	5	2	1	5	3	1
土肥実平	14	9	6	5	0	1	4	4	3	0
以仁王	11	3	0	2	1	0	0	2	1	5
千葉常胤	10	2	10	1	4	0	0	3	4	1
佐々木定綱	10	6	3	5	4	2	3	0	4	3
源義朝	9	4	5	3	3	1	4	4	0	2
平清盛	9	3	1	1	0	5	1	3	2	0

次に，抽出された人名の中から注目した特定人物と他の人物との共起頻度を算出する．人物間の共起頻度を算出するために，テキストマイニングの前処理を行う TinyTextMiner[6]というソフトウェアを用いて処理を行う．

本手法では，日記体の古典史料を対象としているため，同じ日付の日記に登場する人物同士を共起しているとし，この共起を全日付において累計することにより，人物間の共起頻度を算出する．表 3 は，算出された人物間の共起頻度情報の一部を示す．上記で得られた共起頻度の統計情報を用いることにより，注目した特定人物と他の人物との関係を導く．

人物関係を導くために，統計解析ソフトウェアである R[7]を用いて散布図を作成する．R ではグラフ表示したい共起頻度情報を読み込み，読み込んだ値を散布図として表示する．図 1 は，提案手法による現代語訳の解析から人物同士の関係を導く流れを示す．

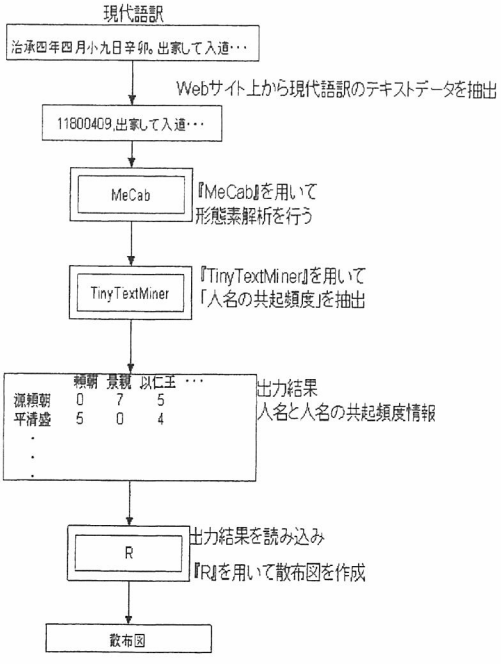

図 1：人物同士の関係を導く流れ

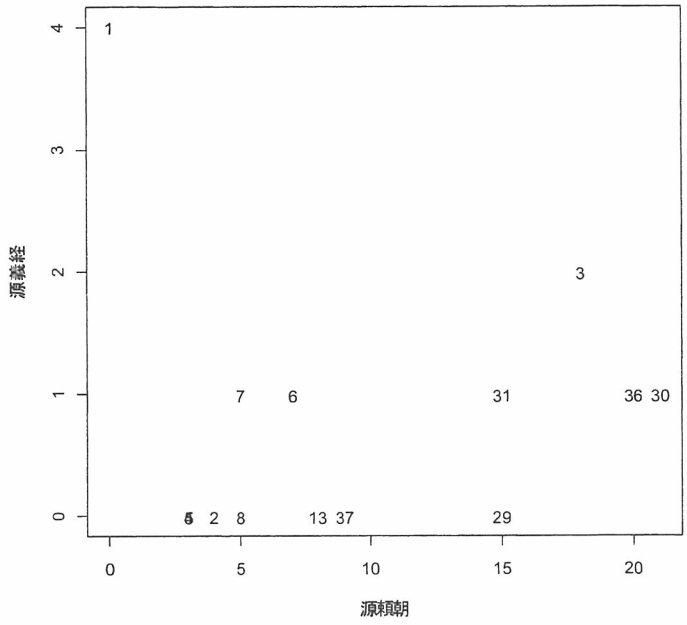

図 2：1184 年の人物関係

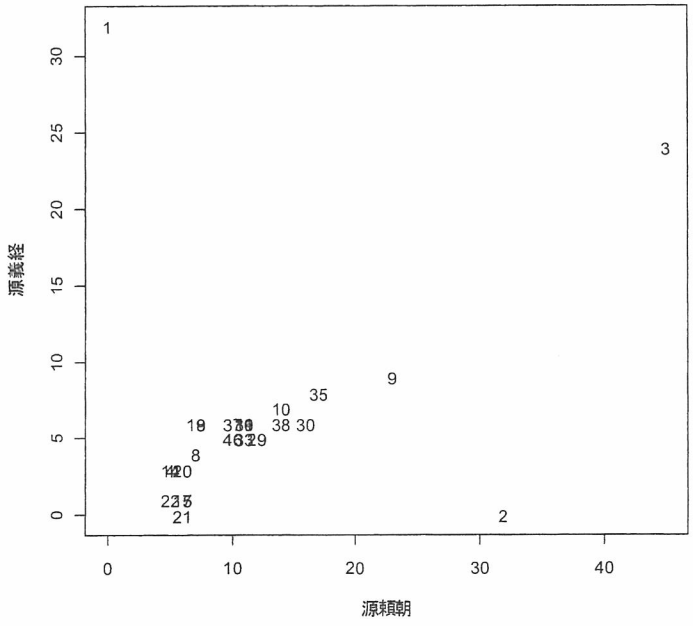

図 3：1185 年の人物関係

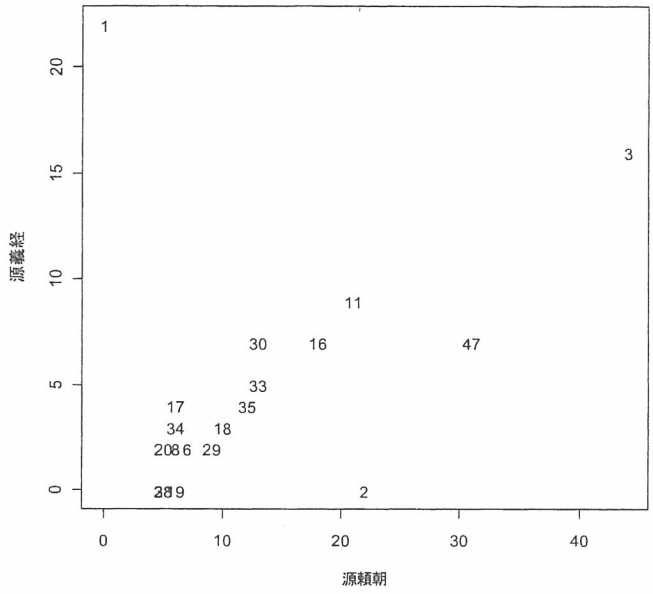

図 4：1186 年の人物関係

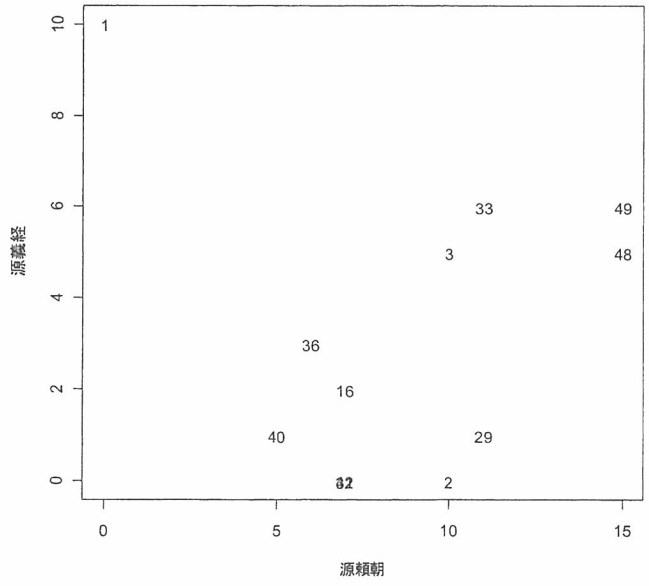

図 5：1187 年の人物関係

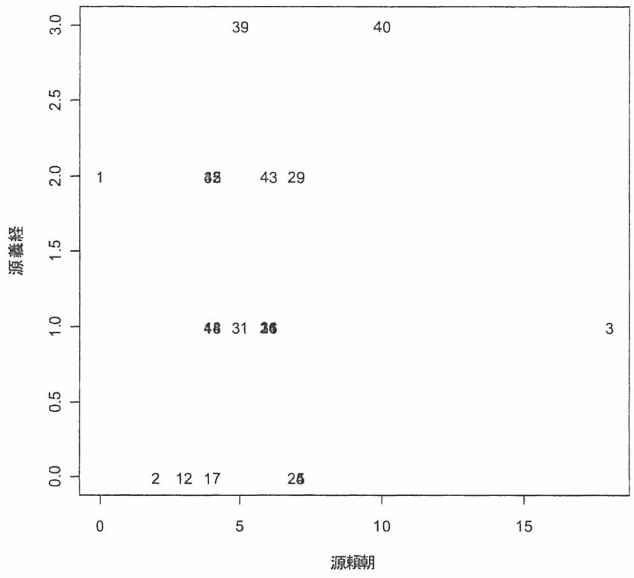

図 6：1188 年の人物関係

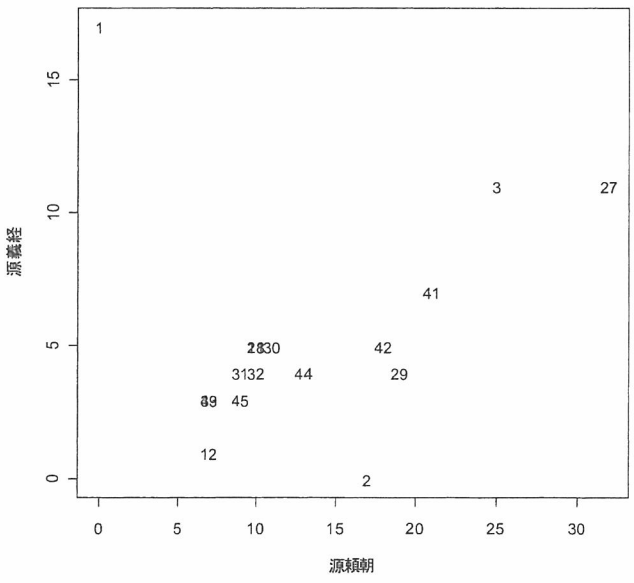

図 7：1189 年の人物関係

表 4：人物 ID 表

ID	Name	ID	Name	ID	Name	ID	Name
1	源頼朝	14	平時忠	27	藤原泰衡	40	諏方盛澄
2	源義経	15	平宗盛	28	藤原朝光	41	宇佐美祐茂
3	後白河天皇	16	北条時政	29	千葉常胤	42	吉見頼綱
4	以仁王	17	北条政子	30	糟谷有季	43	吉香友兼
5	平清盛	18	藤原親経	31	小山朝政	44	梶原景時
6	源頼政	19	藤原忠親	32	小山朝光	45	三浦義澄
7	源義広	20	藤原雅長	33	佐佐木定綱	46	孝謙天皇
8	源義朝	21	藤原能成	34	佐佐木盛綱	47	六条院
9	源範頼	22	藤原親光	35	原田種直	48	白河天皇
10	源師仲	23	藤原実定	36	熊谷直実	49	天武天皇
11	源通親	24	藤原宗長	37	比企能員		
12	源頼家	25	藤原能盛	38	岡辺泰綱		
13	平敦盛	26	藤原任憲	39	新田義兼		

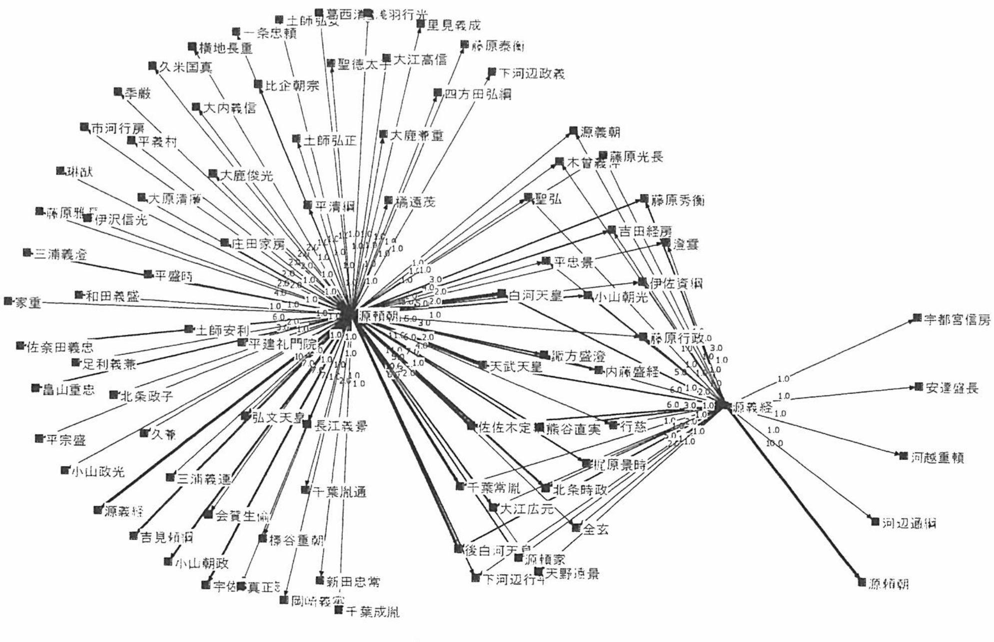

図 8：相関関係ネットワーク

5. 出力結果

1184 年から 1189 年の年代別に源頼朝と源義経に共起する人物を散布図にプロットして関係を表した．散布図にプロットされている数字は，人物を表す ID である．これは，散布図に人名をプロットした際に，人名がプロットされると文字幅があるため，正確な位置が把握できないことと人名が重なって表示され分かりにくくなることを防ぐために，人物を表す ID の作成を行った．図 2 に 1184 年,図 3 に 1185 年,図 4 に 1186 年,図 5 に 1187 年，図 6 に 1188 年，図 7 に 1189 年を，表 4 に散布図にプロットする人物を表す人物 ID 表を示す．

6. 考察

図 2 と図 3 を比較すると，図 2 では源頼朝と共起する人物が多かったが，図 3 では源義経と共起する人物が増え，後白河法皇との共起が大幅に増えている．これは，源義経が後白河法皇（ID=3）より判官の役職を授かったため多くの人物と関係を持つようになったと考えられる．図 3 と図 4 を比較すると，大きな変化はないが，図 4 では源頼朝と共起する人物が増え，源義経と共起する人物が減っている．今後，源頼朝が源義経を追討するために，勢力が変化していると考えられる．図 4 と図 5 を比較すると，図 4 では源義経と共起する人物が多くいたが，図 5 では共起する人物が減り，源頼朝と共起する人物が多いことが目立つ．これは，源頼朝が人々に源義経に従わないよう命令したことが反映されおり，源頼朝と源義経の関係が悪化していることが分かる．図 5 と図 6 を比較すると，源頼朝と源義経の関係悪化により，源義経と共起する人物が少なくっていると考えられる．1189 年は，源頼朝が藤原泰衡を討った年のため，多くの人物と繋がりがあり，藤原泰衡を共起が多いことが表されている．

共起頻度情報を基に，注目した人物（今回の実験では，源頼朝と源義経）とその他の人物の関係を年代別に散布図で表すことにより，注目した人物の勢力の推移や繋がりの変化が可視化される．

7. あとがき

本論文では，吾妻鏡の現代語訳を対象として形態素解析を行い，人名に注目して抽出し，特定人物とその他の人物関係を散布図によって導いた．今回は，『吾妻鏡』の 1184 年から 1189 年の源頼朝と源義経に共起する人物を散布図にプロットを行い，2 人に共起する人物の変化を表した．今回と同様の手法を適用し，違う年代の日記に対して，注目した人物とその他の人物との変化や，源氏と平氏の勢力の推移を導くことができると考える．同様の手法により，人名だけでなく地名なども抽出することにより，人物と地名と年代の関係を導き，3 次元で表示し，比較することで新たな知識を得ることを考えている．

今後の課題として，様々な可視化方法を提案し，解析結果を表示することで，古典史料から知識獲得を行う手法の確立を目指す．例えば，図8の相関関係ネットワークである．相関関係ネットワークは，UCINET[9]を用いて作成する．相関関係ネットワークでは，人物間の繋がりを線の太さで表すことができる．人物間の繋がりの強度を線の太さだけでなく，注目する人物からの距離で強さを表す方法も考えている．

さらに，現状では古典史料の現代語訳があるものを対象として解析することにより知識獲得を行っているが，将来的には古典史料の原文を直接テキストマイニングによって解析を行うことを検討している．古典史料の原文を直接解析することにより，新たな歴史の発見に繋がることに期待される．

8. 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] 村川 猛彦，山中 克真，宇都宮 啓吾：古典籍書誌情報におけるキーワード抽出手法，情報知識学会誌，Vol.18，No.2，pp.87-92（2008）
[2] 歴散加藤塾：吾妻鏡入門，http://katohjuk-okuji.html
[3] 市村 由美，長谷川 隆明，渡部 勇，佐藤 光弘：テキストマイニング 事例紹介，人工知能学会誌，pp.192-200（2001）
[4] 工藤 拓：MeCab，http://mecab.sourceforge.net/
[5] 吾妻鏡データベース：吉川弘文館（2002）
[6] 松村真宏，三浦麻子：TTM:TinyTextMiner β version，http://mtmr.jp/ttm/
[7] 熊谷 悦生，舟尾 暢男：R で学ぶデータマイニング データ解析の視点から，pp.91-105，九天社（2007）
[8] AnalyticTechnokogies：UCINET，http://www.analytictech.com/ucinet/