

# Wikipedia からの法則の変遷情報の抽出

範 良良<sup>\*1</sup>      村田 真樹<sup>\*1</sup>      徳久 雅人<sup>\*1</sup>      馬 青<sup>\*2</sup>

<sup>\*1</sup> 鳥取大学 大学院 工学研究科 情報エレクトロニクス専攻

<sup>\*2</sup> 龍谷大学 理工学部 数理情報学科

<sup>\*1</sup>{k112001,murata,tokuhisa}@ike.tottori-u.ac.jp

<sup>\*2</sup>qma@math.ryukoku.ac.jp

## 1 はじめに

近年 Wikipedia というオンライン百科事典が世界中で広く利用されている。Wikipedia から、どの法則からどの法則が生まれたかといった法則の変遷情報を抽出しユーザに提示できれば便利である<sup>1</sup>。

そこで、本研究では法則の変遷情報を Wikipedia から抽出することを目指す。本論文での変遷情報は、変遷の関係にある法則対と各法則の発見年 (以降、法則年号と呼ぶ) と定義する。変遷情報の例を表 1 に示す。

本研究での変遷情報の抽出方法は、以下のとおりである。法則ページ (法則を記載したページ) に記載されている年号より各法則の発見年を予測し、ある法則 A のページに他の法則 B が記載されている場合に法則 A と法則 B が変遷の関係にある可能性が高いとするヒューリスティックルールに基づき、法則 A と法則 B の対をそれぞれの法則の発見年とともに変遷情報として抽出する。法則年号の抽出は、法則ページの先頭の年号が法則年号になりやすいというヒューリスティックルールに基づいて行う。ヒューリスティックに基づく手法の性能を向上させるために、教師あり機械学習を用いて性能を向上させることも行う。

本論文の主な主張点は以下のとおりである。

- 法則ページの先頭の年号を法則年号とし、基本法則と関係法則の対を変遷情報として取り出すというヒューリスティックルールに基づく手法を提案した。この簡単な手法でも F 値 0.46 で変遷情報を取得できた。
- 上記のヒューリスティックルールに加え教師あり機械学習法を利用する手法を提案した。この手法により性能を改善させ、F 値 0.68 で変遷情報を取得できた。

<sup>1</sup>法則の変遷情報の抽出の意義には以下のものがある。法則の変遷情報は、法則の基本的な情報であり、収集し整理できると便利である。法則間の関係をより理解しやすくなる。また、堀らの研究 [1] のように、科学の発展の歴史を整理することにも役立つ。

表 1: 変遷情報の例

法則 A	法則 B
SMILES 記法 (1980 年)	グラフ理論 (1736 年)
ゲーム理論 (1928 年)	決定理論 (1670 年)

- 変遷の関係にある法則対の取り出しでは (法則年号は取り出さなくてよい)、教師あり機械学習法を利用することで 0.87 という高い F 値を得た。
- 提案手法は、本課題と同様な構成を取る問題に適用することができる。例えば、Wikipedia にある、年号を持つ他の種類のページ群からそのページ群に関わる変遷情報を取得することに応用できる。

## 2 関連研究

堀ら [1] は情報抽出の研究として、研究者および研究分野の情報を重み付け手法で分析することで、研究者および研究分野の変遷情報を自動的に抽出した。隅田ら [2] は Wikipedia の記事構造に含まれる節や箇条書きの見出しから、大量の上位下位関係候補を抽出し、機械学習を用いてフィルタリングすることで高精度の上位下位関係を獲得する手法を開発した。新井ら [3] は連想シソーラスの構築を目指し、手法としてリンク共起性解析を利用した。

## 3 提案手法

変遷情報の抽出手法としてヒューリスティックルールに基づく手法と教師あり機械学習に基づく手法を提案する。教師あり機械学習には性能の優れたサポートベクターマシン (SVM) を利用する (カーネル関数には 2 次の多項式カーネルを利用する)。

表 2: 手法 A2 の素性

利用した素性	内容
f1	年号前後の文字列
f2	文頭から年号までの文の長さ

変遷情報の抽出は提案手法に基づき、法則年号の抽出と法則対の抽出により行う。法則年号の抽出と法則対の抽出のそれぞれの詳細な手法について以下に述べる。

### 3.1 法則年号の抽出

Wikipedia の法則ページから法則の発見年を抽出する。Wikipedia の法則ページに法則の発見年は記載されている場合が多い。これを利用し法則ページから法則年号の抽出を行う。

法則ページから法則年号を抽出するための 3 つの手法を下記に示す。手法 A1 はヒューリスティックルールに基づく手法であり、手法 A2 と手法 A3 は教師あり機械学習に基づく手法である。

**手法 A1** 法則ページの最初の年号をその法則の発見年として出力する手法。法則ページの最初に出現した年号は法則の発見年である場合が多いことから、その最初の年号を抽出し法則年号とする。このとき、抽出した法則年号はこの手法の出力になる。

**手法 A2** 法則ページの最初の年号を取り出し、その年号は法則の発見年であるかどうかを機械で判断する手法。手法 A1 と異なり、手法 A2 の場合は機械の判断により抽出した年号は法則の発見年でない場合は出力はしないものとし、法則の発見年である場合はその年号を出力とする。

**手法 A3** 法則ページの全部の年号を取り出し、取り出した全部の年号を機械学習 SVM によって評価しスコアをつけ、スコアが最も高い年号を出力とする。スコアの最も高い年号のスコアが負(年号が正しくないを意味する)の場合は、出力はしないものとする。

法則年号の抽出の教師あり機械学習に基づく手法(手法 A2 と手法 A3)の素性として利用したものを表 2、表 3 に示す。

抽出した年号が西暦でない場合は西暦変換を行う。西暦変換とは、法則年号が西暦であるかどうかをチェックし、西暦でない場合その法則年号をプログラムによって西暦に変換する。西暦変換の例を表 4 に示す。

表 3: 手法 A3 の素性

利用した素性	内容
f1	年号前後の文字列
f2	文頭から年号までの文の長さ
f3	年号の順番

表 4: 西暦変換の例

	西暦
昭和 34 年	1959 年
紀元前 1000 年	-1000 年

### 3.2 法則対の抽出

本研究では法則ページのタイトルとなる法則を基本法則と呼び、法則ページに存在する他の法則のことを関係法則と呼ぶ。変遷の関係にある法則対をルーツ法則と派生法則と呼び、そのとき法則年号の早い方はルーツ法則になる。これを表 1 の例から説明すると、法則対 “SMILES 記法 (1980 年) グラフ理論 (1736 年)” は変遷情報であるため、ルーツ法則と派生法則の対になる。このとき法則 “グラフ理論” の発見年 (1736 年) が法則 “SMILES 記法” の発見年 (1980 年) より早いため、法則 “グラフ理論” はルーツ法則になる。

法則ページから抽出した基本法則と関係法則の対には変遷の関係である法則対が多い。そのため、法則対の抽出では基本法則と関係法則の対から変遷の関係である法則対、すなわちルーツ法則と派生法則の対を抽出する。

法則ページからルーツ法則と派生法則の対を抽出するための 2 つの手法を下記に示す。手法 B1 はヒューリスティックルールに基づく手法であり、手法 B2 は教師あり機械学習に基づく手法である。

**手法 B1** 法則ページから取り出した基本法則と関係法則の対すべてを変遷の関係であると判断する手法。

**手法 B2** 法則ページから取り出した基本法則と関係法則の対が変遷の関係であるかどうかを機械で判断する手法。機械の判断により抽出した法則対が変遷の関係でない場合は出力をしないものとし、変遷の関係である場合はその法則対を出力とする。

法則対の抽出において教師あり機械学習に基づく手法(手法 B2)の素性として利用したものを表 5 に示す。

表 5 の双方向法則対を以下で定義する。ある法則 C と法則 D の対に対し、もし法則 C のページに法則 D

表 5: 手法 B2 の素性

利用した素性	内容
f1	法則対の名前類似度
f2	法則対は双方向法則対であるかどうか

表 6: 双方向法則対の例

基本法則	関係法則
法則 C : ゲーム理論	法則 D : 決定理論
法則 D : 決定理論	法則 C : ゲーム理論

が記載されており、かつ逆に法則 D のページに法則 C も記載されている場合、この法則対を双方向法則対と呼ぶ。双方向法則対の例を表 6 に示す。

### 3.3 変遷情報の抽出

法則対の抽出で取り出した変遷の関係にある法則対を、法則年号の抽出で取り出した法則の発見年とともに抽出することで、変遷情報を抽出する。変遷情報の抽出の手法は、法則年号の抽出の 3 つの手法と法則対の抽出の 2 つの手法を組み合わせることにより以下の 6 つの手法になる。

手法 C1 手法 A1 と手法 B1 を利用する。

手法 C2 手法 A2 と手法 B1 を利用する。

手法 C3 手法 A3 と手法 B1 を利用する。

手法 C4 手法 A1 と手法 B2 を利用する。

手法 C5 手法 A2 と手法 B2 を利用する。

手法 C6 手法 A3 と手法 B2 を利用する。

## 4 実験

### 4.1 実験データ

実験には、Wikipedia から 2010 年 5 月 26 日にダウンロードしたページを利用する。そのデータの内訳は以下のとおりである。法則ページは 1,634 個である。抽出した法則対は、延べで 2,074 個、異なりで 1,621 個である。法則ページの取り出しはパターンにより取り出した後人手でチェックして行った。

1,621 個の異なる法則対からランダムに取り出した 100 個の法則対をデータ A(異なる法則 : 133 個)、デー

表 7: 法則年号の抽出の結果

手法	再現率	適合率	F 値
手法 A1	0.92	0.61	0.74
手法 A2	0.76	0.85	0.80
手法 A3	0.68	0.83	0.75

表 8: 法則対の抽出の結果

手法	再現率	適合率	F 値
手法 B1	1.00	0.60	0.75
手法 B2	0.87	0.88	0.87

タ A と重複せずにランダムに取り出した他の 100 個の法則対をデータ B(異なる法則 : 137 個) とする。

### 4.2 性能の計算

本実験で抽出の性能を測るための再現率、適合率、F 値を以下のとおりに定義する。

$$\text{再現率} = \frac{\text{手法の出力のうちの正解数}}{\text{実際の正解の数}} \quad (1)$$

$$\text{適合率} = \frac{\text{手法の出力のうちの正解数}}{\text{手法による出力の数}} \quad (2)$$

$$F \text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \quad (3)$$

### 4.3 法則年号の抽出

教師あり機械学習の実験では、データ A の 133 個の法則を学習データ、データ B の 137 個の法則をテストデータとする。実験の結果を表 7 に示す。

### 4.4 法則対の抽出

教師あり機械学習の実験では、データ A の 100 個の法則対を学習データ、データ B の 100 個の法則対をテストデータとする。実験の結果を表 8 に示す。

### 4.5 変遷情報の抽出

変遷情報の抽出精度を評価するために 3 つの基準を定めた。

**基準 1** 法則対の出力に対して，法則対が実際に変遷の関係であり，かつ法則対の法則年号が正しい場合，正解と判断する基準。

**基準 2** 法則対の出力に対して，法則対が実際に変遷の関係であり，法則年号が間違っている場合，間違った年号により正しい順番が推定できる場合，正解と判断する基準。

**基準 3** 法則対の出力に対して，法則対が実際に変遷の関係である場合，正解と判断する基準（即ち，法則年号は取り出さなくてよい）。

その 3 つ基準を変遷情報の抽出の各手法に適用した結果を表 9 に示す。

## 5 考察

表 7 と表 8 より，法則年号，変遷の関係にある法則対ともに，どの手法でも 0.7 から 0.8 という高い F 値を獲得できることがわかる。特に，変遷の関係にある法則対は，機械学習を使うことで，0.87 というかなり高い F 値を得た。

次に表 9 により，変遷情報の抽出の性能を考察した。ヒューリスティックルールに基づく手法（手法 C1）で，F 値 0.46 を得た。法則ページの先頭の年号を法則年号とし，基本法則と関係法則の対を変遷情報として取り出すというヒューリスティックルールだけでも，この性能を得られることがわかった。

教師あり機械学習法を利用することで F 値 0.68 で変遷情報を取得できた。上記のヒューリスティックルールに加え教師あり機械学習法（手法 C2）を利用することで性能の改善が可能であることがわかった。

基準 3 の変遷の関係にある法則対の取り出しでは（法則年号は取り出さなくてよい），教師あり機械学習法（手法 C4,C5,C6）を利用することで 0.87 という高い F 値を得た。

## 6 おわりに

本研究では Wikipedia から法則の変遷情報を抽出する手法を提案した。ヒューリスティックルールと教師あり機械学習に基づく手法を用いて抽出を行った。ヒューリスティックルールは，法則ページの先頭の年号を法則年号とし，基本法則と関係法則の対を変遷情報として取り出すというものである。実験の結果，変遷情報の抽出ではヒューリスティックルールに基づく簡単な手法でも F 値 0.46 を得た。ヒューリスティ

表 9: 変遷情報の抽出の結果

手法	基準	再現率	適合率	F 値
C1	1	0.97(30/31)	0.30(30/100)	0.46
	2	0.97(30/31)	0.30(30/100)	0.46
	3	1.00(60/60)	0.60(60/100)	0.75
C2	1	0.71(22/31)	0.65(22/ 34)	0.68
	2	0.71(22/31)	0.65(22/ 34)	0.68
	3	1.00(60/60)	0.60(60/100)	0.75
C3	1	0.48(15/31)	0.60(15/ 25)	0.54
	2	0.52(16/31)	0.64(16/ 25)	0.57
	3	1.00(60/60)	0.60(60/100)	0.75
C4	1	0.71(24/34)	0.41(24/ 59)	0.52
	2	0.71(24/34)	0.41(24/ 59)	0.52
	3	0.87(52/60)	0.88(52/ 59)	0.87
C5	1	0.50(17/34)	0.71(17/ 24)	0.59
	2	0.50(17/34)	0.71(17/ 24)	0.59
	3	0.87(52/60)	0.88(52/ 59)	0.87
C6	1	0.35(12/34)	0.71(12/ 17)	0.47
	2	0.35(12/34)	0.71(12/ 17)	0.47
	3	0.87(52/60)	0.88(52/ 59)	0.87

クルールに用いた情報を素性として利用した教師あり機械学習に基づく手法で F 値 0.68 を得た。法則年号を取り出さなくてよく，変遷の関係にある法則対を取り出すという目的では，教師あり機械学習手法で F 値 0.87 を得た。

## 参考文献

- [1] 堀さな子，村田真樹，徳久雅人，馬青：“研究者および研究分野の変遷の自動推定”，言語処理学会第 17 回年次大会発表論文集，pp.236–239，2011.
- [2] 隅田飛鳥，吉永直樹，鳥澤健太郎：“Wikipedia の記事構造からの上位下位関係抽出”，自然言語処理，16(3)，pp.3–24，2009.
- [3] 新井嘉章，福原知宏，増田英孝，中川裕志：“Wikipedia からの連想シソーラス構築プロジェクト”，第 20 回セマンティックウェブとオントロジー研究会，SIG-SWO-A803-15，2009.
- [4] Wikipedia: <http://ja.wikipedia.org/wiki/>