概要

Wikipedia というオンライン百科事典には，多くの法則が収録されている．法則には別の法則から発見されるという変遷があり，読者は法則の変遷を知ることで法則の理解を深めることができる．そこで，本研究では法則の変遷についての情報を Wikipedia から自動抽出することを目的とする．

本研究での変遷情報とは，変遷の関係にある 2 つの法則名および各発見年による組と定義する．たとえば，「1670 年に発見された決定理論を基に，1928 年にゲーム理論が提唱された」という文では，『「決定理論 (1670)」「ゲーム理論 (1928)」』という組を 1 つの変遷情報とする．

変遷情報の抽出は法則年号の抽出処理と法則対の抽出の 2 つで構成する．変遷情報の抽出手法として，ヒューリスティックルール，および，教師あり機械学習に基づく手法を提案する．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．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．ヒューリスティックルールに加え教師あり機械学習を利用する手法で 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．

目 次

図 目 次

表 目 次

第1章　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おり，Wikipedia に関する様々な研究がなされている．Wikipedia というオンライン百科辞典には，多くの法則が収録されている．法則には別の法則から発見されるという変遷があり，読者は法則の変遷を知ることで法則の理解を深めることができる．そこで，本研究では法則の変遷についての情報を Wikipedia から自動抽出することを目的とする．

本研究での変遷情報とは，変遷の関係にある 2 つの法則名および各発見年による組と定義する．たとえば，「1670 年に発見された決定理論を基に，1928 年にゲーム理論が提唱された」という文では，『「決定理論 (1670)」「ゲーム理論 (1928)」』という組を 1 つの変遷情報とする．変遷の関係である法則対と変遷の関係でない法則対を一つずつ表 1.1 に示す．

表 1.1: 変遷情報の候補

法則 A	法則 B	結果
決定理論 (1670 年)	ゲーム理論 (1928 年)	変遷情報
擬似乱数 (1948 年)	制御理論 (1950 年)	非変遷情報

変遷情報の抽出は法則年号の抽出処理と法則対の抽出の 2 つで構成する．変遷情報の抽出は以下のとおりに行う．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．変遷情報の抽出手法として，ヒューリスティックルール，および，教師あり機械学習に基づく手法を提案する．なお，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．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．

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法則間の関係をより理解しやすくなる．また，科学の発展の歴史を整理することにも役立つ．

本研究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，基本法則と関連法則 (3.1.1 節) 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．
- 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応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．

本論文の構成は以下の通りである．第 2 章では関連研究を説明する．第 3 章では，提案手法，抽出の手順について述べる．第 4 章では，実験データおよび実験の結果について述べる．第 5 章では，機械学習で利用した素性について分析を行う．第 6 章では，Wikipedia における法則以外の変遷に対する調査を行う．第 7 章では，今後の課題を挙げる．第 8 章では，全体をまとめる．

第2章　関連研究

本章では，変遷情報の抽出に関する先行研究，および Wikipedia からの情報抽出に関する研究を説明する．

2.1　変遷情報の抽出に関する研究

変遷情報の抽出に関する研究として，堀らの研究があげられる [1][2]．堀らは研究者や研究分野の変遷情報 (例えば，人名では「池原悟 (先輩) →村上仁一 (後輩)」のような先輩後輩関係の対，分野名では「情報抽出 (ルーツ) →要約 (派生分野)」のような派生関係の対) を自動的に抽出する方法を提案した [1]．論文の著者として，ある人名 A が出現した最初の時期に同時に共起し，人名 A より初出現年が早い人名 B は, 人名 A のルーツ (先輩) である可能性が高いと思われる．また分野名においても同様のことがいえる．この仮説に基づいて研究者と研究分野の変遷情報を抽出した．この手法で抽出した人名の変遷情報，分野名の変遷情報を表 2.1，表 2.2 に示す．しかし，この研究の手法は学術分野間，師弟間という限定された種類の変遷情報しか抽出することができない．

表 2.1: 人名の変遷

人名 A(後輩)	人名 B(先輩)
村上仁一	池原悟
馬青	井佐原均
宮尾祐介	辻井潤一
丸山岳彦	柏岡秀紀
黒田航	井佐原均

より多くの種類の変遷情報を自動で，より高性能に取得することを目的として，堀らはパターンに基づく手法と機械学習を組み合わせることで，大量の文から幅広い変遷情報を取得した [2]．また，抽出した変遷情報は，様々な種類の情報が混ざっているため，変遷情報の自動分類を行った．変遷情報の抽出と分類は以下の手順で行う．

表 2.2: 分野名の変遷

分野名 A	分野名 B(ルーツ)
自動評価	機械翻訳
統計的機械翻訳	統計
情報分析	分析
言語横断情報検索	情報検索
論文要約	情報抽出

1. 大量の文から人手で作成したパターンを利用し，変遷情報を自動で抽出する．また，教師あり機械学習を追加してより高性能に変遷情報の抽出を行う．

2. 1 で抽出した変遷情報は何についての変遷かわからないため，1 で抽出した変遷情報を人手で分類し，分析する．

3. 機械学習を利用して変遷情報の自動分類も行う．

変遷情報の自動分類は変遷情報の含み方、変遷の種類、変化の仕方に基づいて行った．ここで，変遷情報の含み方に基づく分類について説明する．変遷情報の含み方に基づく分類の例を図 2.1 に示す．

type-A X，Y が明らかに変遷情報であり，知見の得られる事例．ただし，X，Y 自体が変遷関係にない場合であっても, X,Y に対して修飾関係 (連接した修飾関係) にある語が変遷関係にある場合も type-A とする．

type-B X，Y のどちらか一方が一般的に広い意味を持つ名詞であるが，文の構造からその名詞の具体的内容を示す表現がその文の他の個所から抽出できる事例．

type-C X，Y のどちらか一方が一般的に広い意味を持つ名詞であるが，X，Y の名詞から変遷として知見の得られる事例

type-D X，Y のどちらか一方, もしくは両方が一般的に広い意味を持つ名詞であり，変遷として知見の得られない事例

type-E 単に場所を指定している事例

type-F 単に状態を表している事例

分類	文の例
type-A	発生過程を再現するように、ＥＳ細胞を神経 (Y)の元になる幹細胞 (X) などに分化させ、さらに条件を変えて培養することで、前脳型アセチルコリン作動性神経細胞など様々な神経細胞を作り分けることに成功した。 (解説：「幹細胞」は「神経」の元の物質であるため，変遷とみなせる
type-B	精米の目的は、お米の表面近くに分布する、タンパク質や粗脂肪などのお酒の雑味 (Y)の元となる成分 (X) を取り除くことにあります。 (解説：「成分」は一般的に広い意味を持つ名詞であるが「タンパク質や粗脂肪などの」で修飾されている)
type-C	臭いや にきび (Y)の元となる原因菌 (X) の殺菌効果に優れたボディソープです。 (解説：「原因菌」は一般的に広い意味を持つ名詞であるが X と Y を 2 つ見て知見が得られる)
type-D	ＪＡＡＡ３０年の歴史は、まさに「自動化に絡む 企業活動 (X)から派生する関係性 (Y) を楽しむ充足感」を動機として運営されてきた (解説：「関係性」は一般的に広い意味を持つ名詞であり，知見が得られない．)
type-E	また同社はブルゴーニュ全域でも最大の土地所有者のひとつに数えられ、１００ｈａの自社畑 (X)から生まれるワイン (Y) は、同社の生産量の８５％を占める。 (解説：「自社畑」は場所を示している．)
type-F	同じような境遇 (X)で生まれた組織の先輩 (Y) には『００７美しき獲物たち』の悪役、マックス・ゾリンがいます。 (解説：単に「組織の先輩」の状態を表している)

図 2.1: 変遷情報の含み方に基づく分類ごとの文の例 (文献 [2] より引用)

堀らの研究と本研究との比較を以下にまとめる．

1. 変遷情報の抽出という点では，堀らの研究と本研究は類似しているが，堀らの研究は，研究者や研究分野の変遷情報を抽出し，さらに文章中から変遷情報を抽出ことを目的とする．それに対し，本研究はWikipediaから法則の変遷情報の獲得を目的とする．

2. 研究の手法として，堀ら [1] は重み付け手法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．堀ら [2] はパターンに基づく手法と機械学習を組み合わせることで，文章中から変遷情報を取得した．本研究は変遷情報の抽出手法としてWebのリンク情報に基づくヒューリスティックルールと教師あり機械学習を組み合わせる手法を提案する．

3. 堀らの手法は文章中の変遷を示す表現を用いて変遷情報を抽出するが，本研究の手法はWikipediaのリンク情報を利用して法則の変遷情報を抽出する．

2.2 Wikipediaからの情報抽出に関する研究

2.2.1 Wikipediaの記事構造からの上位下位関係抽出

隅田らは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で上位下位関係を獲得する手法を開発した [3]．

ここで，隅田らの研究手法，特に本研究の参考になった素性の設定について説明する．まず，隅田らの研究手法について説明する．隅田らの研究手法は以下のとおりである．

Step1 Wikipediaの記事構造からの上位下位関係候補の抽出 このステップでは，記事構造の各ノードを上位語候補，子孫関係にあるノードを下位語候補とする全ての組み合わせを上位下位関係候補として抽出する．例えば，図 2.2 の記事構造からは，「ブレンドティー/チャイ」や,「紅茶/リプトン」などの上位下位関係候補が抽出できる．

Step2 機械学習によるフィルタリング Step1 の手続きで得られた上位下位関係候補は多くの適切な関係を含む一方で，「生産地/インド」「紅茶ブランド/イギリス」のような誤りも含む．Step2 では，Step1 で抽出した上位下位関係候補から教師あり

機械学習を用い不適切な関係を取り除く．上位下位関係候補が適切な上位下位関係か否かを判定するため，Support Vector Machine (SVM)(Vapnik 1998) で学習された分類器を用いて上位下位関係候補を選別する．

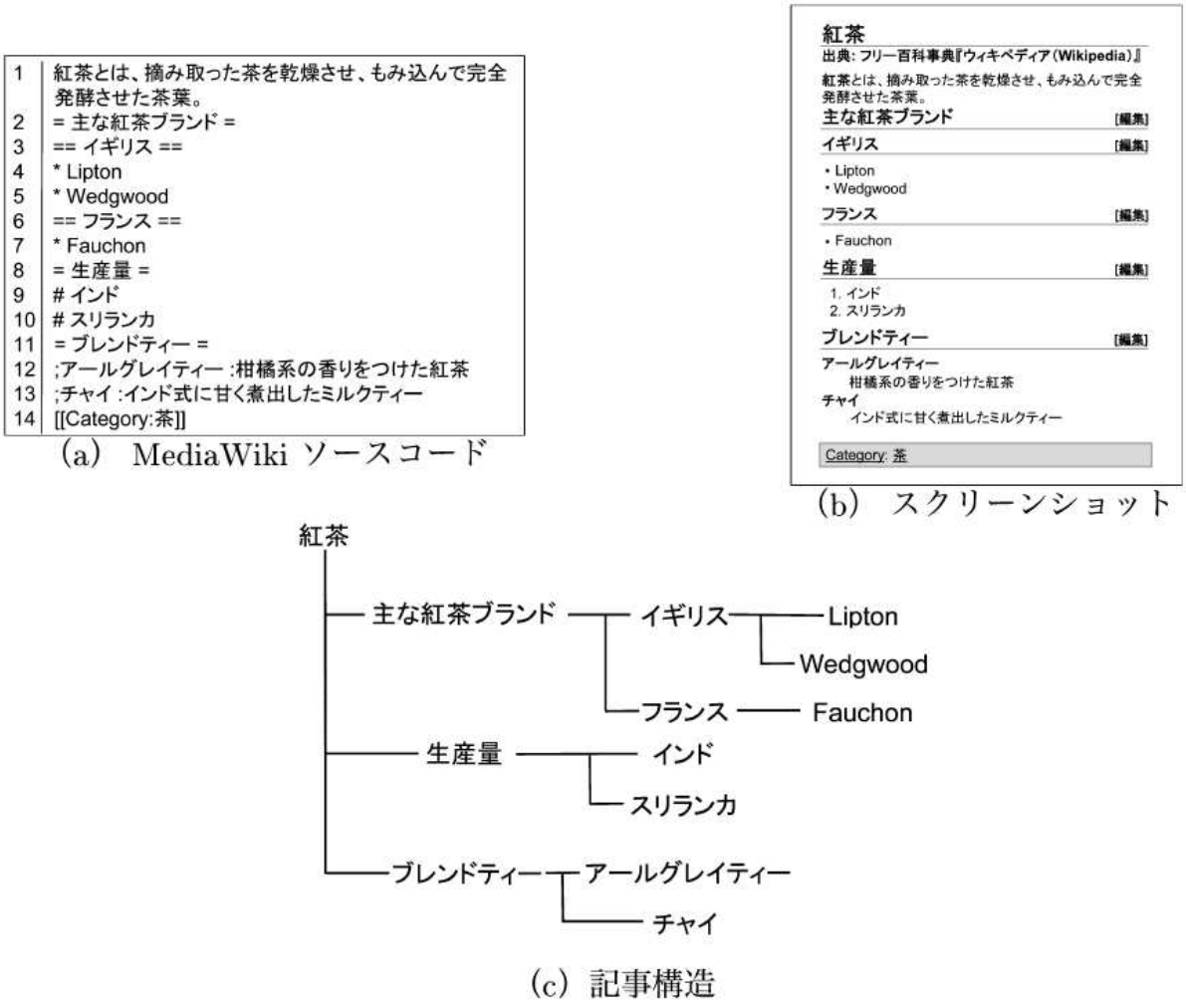

図 2.2: 「紅茶」に関する Wikipedia の記事の例 (文献 [3] より引用)

次に，隅田らの機械学習の実験で用いた素性について説明する．隅田らの研究では素性として，上位下位関係候補がある条件 (特徴) を満たすかどうかを一つの素性として表現し，素性ごとに設定された条件を入力の上位下位関係候補が満たせば，対応する素性ベクトルの次元の値に 1 をセットし，満たさなければ 0 をセットする．

素性として利用するものは,「上位語候補・下位語候補の品詞 (POS)」,「上位語候補・下位語候補中の形態素の表層文字列 (MORPH)」,「不要語 (EXP)」,「属性語 (ATTR)」,「修

飾記号の種類 (LAYER)」,「上位語候補と下位語候補との間の距離 (DIST)」,「子孫ノード (PAT)」,「形態素間の類似性 (LCHAR)」である．以下で，素性「上位語候補と下位語候補との間の距離 (DIST)」と,「形態素間の類似性 (LCHAR)」の設定を説明する．

上位語候補と下位語候補との間の距離 (DIST) 記事構造で上位語候補と下位語候補との間の距離が近ければ近いほど，正しい上位下位関係であることが多い．そこで，記事構造中における上位語候補・下位語候補間の距離を素性とすることで，この傾向を捉える．隅田らの研究では，上位語候補，下位語候補間の距離は記事構造中で上位語候補と下位語候補間に存在する辺の数とする．例えば，図 2.2 の記事構造上で「Wedgwood」と「紅茶ブランド」間の距離は 2 である．素性 DIST では，上位語候補と下位語候補間の距離が 2 以上か否かという 2 つの状態にそれぞれ異なる次元を割りあてた．

形態素間の類似性 (LCHAR) 素性 MORPH では，形態素間の類似性を判断しているため,「高校」や「公立校」のように形態素の一部が一致する語の類似性はないと判断してしまう欠点が存在する．そこで上記のような事例を扱えるようにするため，素性 LCHAR では，上位語候補と下位語候補の末尾の 1 文字が共通する複合語に意味的に似た語が多い特徴を利用し，素性の欠点を補う．具体的には，上位語候補と下位語候補の末尾が同じとき，この MORPH 素性に対応する素性ベクトルの次元の値を 1 にセットするように設計した．

2.2.2 Wikipedia からの連想シソーラス構築プロジェクト

新井らは，リンク共起性解析を用いる手法を提案し，その手法を用いて Wikipedia から大規模で高精度な連想シソーラスを構築した [4]．リンク共起性解析は，ある記事内にある他の記事へのリンクの共起性を解析することによって，リンク先記事が表す概念間の関連度を計算する手法である．本研究では，新井らが提案したリンク共起性解析を参考に，双方向の関係にある法則名のリンクを機械学習 SVM の素性として用いることを提案している．

ここで，新井らが分析した Wikipedia のリンク構造の特徴，および，リンク共起性解析について説明する．まず，新井らが分析した Wikipedia のリンク構造の特徴を以下に説明する．

1. Wikipedia の各記事は，説明のテキスト，図表，そして別の記事に対する多数のリンクで構成される．Wikipedia は Wiki をベースにしており，簡単に他の概念へのリンクを定義できることから，良質な概念どうしのリンクが多いという特徴を持つ．

2. Wikipedia が高密度なリンク構造を持っている．新井らは，予備実験として Wikipedia 内におけるリンク数をカウントしたところ，2006 年 9 月の段階で約 380 万ページ (Redirect リンクを含む) に約 8,000 万の内部リンク (Wikipedia 内へのリンク) を抽出し，Wikipedia では閉じられた語彙空間の中で密なリンク構造を持っているということを確認している．

3. Wikipedia は最新の幅広い分野の記事が網羅されており膨大な量のコンテンツが存在するものの，WWW の探索空間に比較するとそのリンク構造はサイト内で閉じられているため，現実的な時間での解析が可能である．

4. URL によって概念を一意に特定できるという特徴がある．Wikipedia では URL によって一意に示される一つの記事 (ページ) が一つの単語 (概念) を表しており，多義を持つ単語には，意味に応じて別々の記事が用意されている．

次に，リンク共起性解析を説明する．リンク共起性解析は，リンクの共起性を解析することによってリンク間 (記事間) の関連度を算出する．リンクの共起とは，単語をリンクとして扱うということ以外，基本的な概念は単語の共起と同様である．つまり，リンクが共起するということは，特定の範囲においてある異なる二つのリンクが同時に出現するということである．リンクの共起性解析では，リンクは参照先 URL が同じなら同じリンクとみなされ，Wikipedia 全体でのリンクの共起性を解析する．ここで，先に述べたように，Wikipedia におけるリンクは，参照先の記事を一対一で表している．そのため，二つのリンクの関連度を求めることは，Wikipedia の記事が表す二つの概念の関連度を求めることと等価である．

2.3 その他の関連研究

前述した関連研究以外に，以下のような関連研究がある．

山田らは Wikipedia に出現する用語を日本語 WordNet へ追加する手法を提案している [5]．具体的には，Wikipedia におけるタグ情報などを利用して確度高く推定した上位語を利用し，上位語が推定された Wikipedia 中の用語に対して，WordNet の意味概念を表す synset を推定する．

戸田らは，ある特定の興味分野に関する話題の変遷を抽出することにより，話題の変遷が激しいカテゴリ，話題の変遷が緩やかなカテゴリなど，カテゴリごとの話題変遷に関する特徴パタンを抽出する手法を提案している [6].

野田らは，Wikipedia のカテゴリ関係を分析することで，多分野にまたがる意外性のあるカテゴリ関係をもつ項目を発見することを提案した [7].

第3章　提案手法

本章では，変遷情報を自動的に抽出するための手法について説明する．

3.1　提案手法の概要

本研究では，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．

3.1.1　ヒューリスティックルール

ヒューリスティックルールは，法則ページの最初の年号を法則の発見年とし，基本法則と関連法則の対を変遷の関係にある法則対として取り出す手法である．この手法を図 3.1 の例から説明すると，法則「ゲーム理論」のページにおいて，ページの最初に出現した年号 (1928 年) をこの法則の発見年とし，法則ページにリンクの形で記載されている他の法則 (例えば,「決定理論」,「数論」,「集合論」) を「ゲーム理論」と変遷の関係にある法則対 (例えば,『「ゲーム理論」「決定理論」』,『「ゲーム理論」「数論」』,『「ゲーム理論」「集合論」』) として抽出する．

基本法則と関連法則

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連法則と呼ぶ．これを図 3.2 の例から説明すると，「決定理論」という法則のページにおいて，法則ページのタイトルである「決定理論」は基本法則になり，法則ページに存在する他の法則，例えば「ゲーム理論」が関連法則になる．

3.1.2　教師あり機械学習

ヒューリスティックルール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．具体的には，教師あり機械学習を用いてヒュー

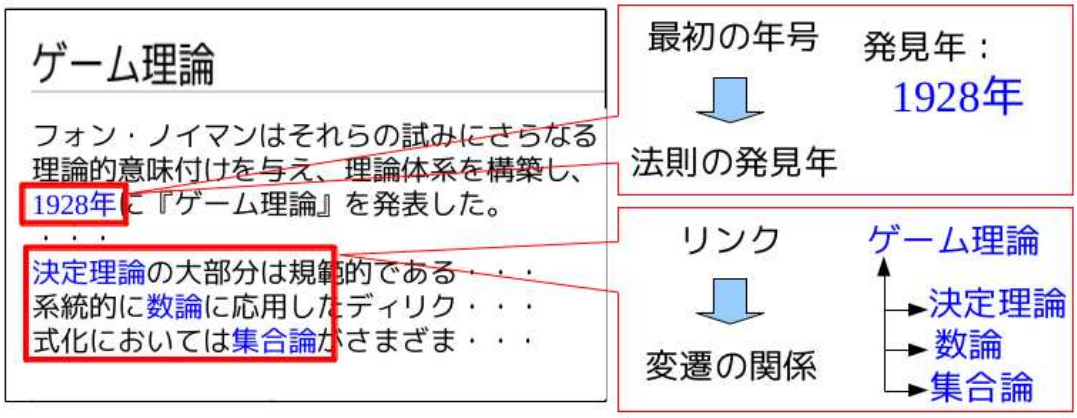

図 3.1: ヒューリスティックルールの例

リスティックルールで取り出した候補が正しいかどうかを判定して，正しいと判定したもののみを取り出す．

3.2 提案手法の詳細

法則の変遷情報の抽出は，法則年号の抽出，法則対の抽出，変遷情報の抽出の順で行う．図 3.3 に示しているように，法則ページに対して，法則年号の抽出処理と法則対の抽出処理を並行して行い，取り出した法則の年号と法則対の情報を合成することで法則の変遷情報を抽出する．本節では，法則年号の抽出，法則対の抽出，変遷情報の抽出で用いる詳細な手法について以下に述べる．

3.2.1 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 の法則ページには法則の発見年が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．法則年号の抽出の例を図 3.4 に示す．

法則ページから法則年号を抽出するための 3 つの手法を以下に示す．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 および手法 A3 は教師あり機械学習に基づく手法である．

図 3.2: 基本法則と関連法則の例

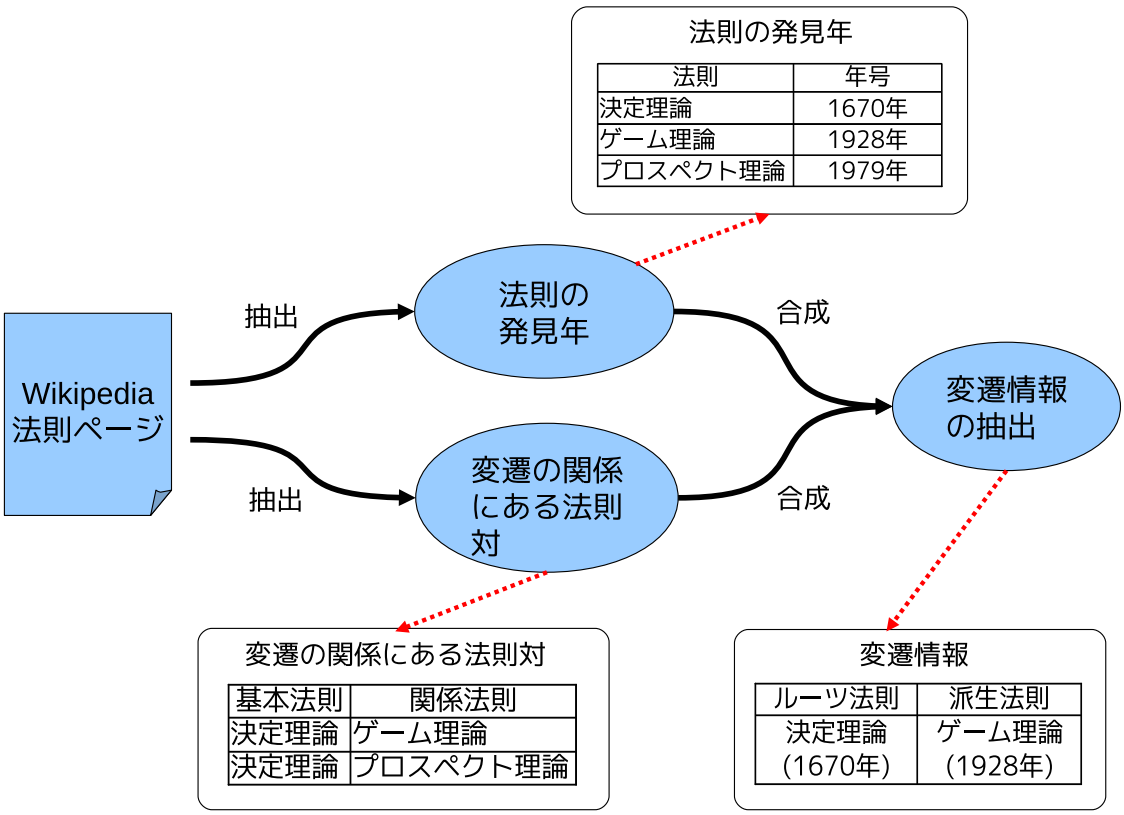

図 3.3: 変遷情報の抽出の流れ

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．

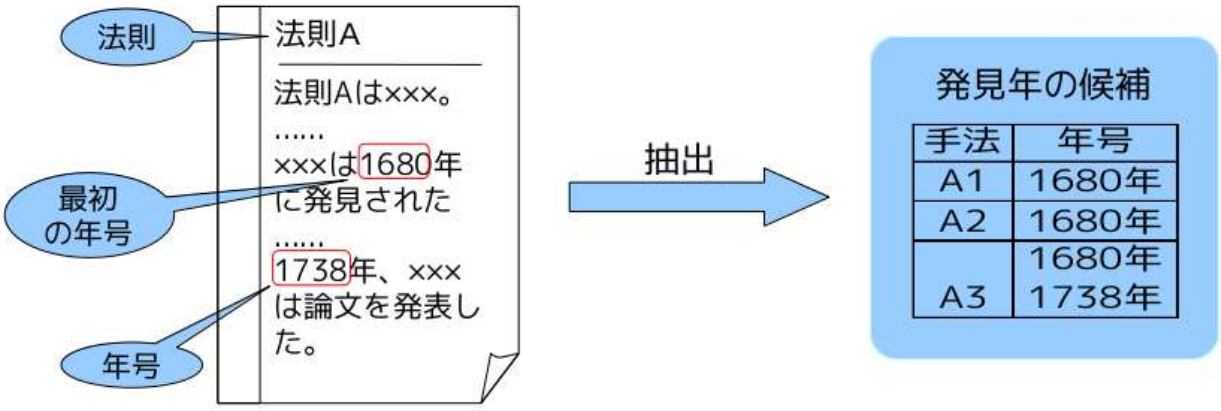

図 3.4: 法則年号の抽出の例

西暦変換

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処理である．西暦変換の例を表 3.1 に示す．例えば，和暦の「昭和 34 年」を西暦変換で行った結果，西暦の「1959 年」になる．

表 3.1: 西暦変換の例

年号	西暦
昭和 34 年	1959 年
紀元前 1000 年	-1000 年

素性の設定

ここで，手法 A2，A3 で利用する素性を表 3.2，表 3.3 に示す．

表 3.2: 手法 A2 で利用した素性

素性 ID	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

表 3.3: 手法 A3 で利用した素性

素性 ID	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

法則年号の抽出で用いる素性は,「年号前後の文字列」,「文頭から年号までの文の長さ」,「年号の順番」である．以下で，この 3 つの素性の設定について説明する．

年号前後の文字列 年号の前と後ろの文字列を利用する．これは，年号の前と後ろにある 5 文字を一文字ずつ削ることで，合わせて 10 通りの表現を生成する．例えば,「周期系に対する DFPT は Baroni らによって 1987 年に提唱された」という文だと，年号 1987 年の前の文字列「らによって」と後ろの文字列「に提唱され」を一文字ずつ削ることで,「らによって」「によって」「よって」「って」「て」「に提唱され」「に提唱さ」「に提唱」「に提」「に」の 10 通りの表現を生成する．

文頭から年号までの文の長さ 文章の先頭から年号が初めて出現した場所までの文の長さを測る．これは，法則の発見年が文頭に出現することが多いという特徴を利用し，文章の先頭から年号が初めて出現した場所までの距離 (文字数) を測る．距離が 1000

文字より小さい場合に，その年号が正解である可能性が高いと考え，この素性に対応する素性ベクトルの次元の値を 1 にセットする．距離が 1000 文字より大きい場合に，その年号が正解である可能性が低いと考え，この素性に対応する素性ベクトルの次元の値を 0 にセットする．

年号の順番 法則ページに出現した年号の順番を利用する．この素性は手法 A3(法則ページの全部の年号を取り出し，機械学習 SVM で判断する手法) で用いる．手法 A3 は法則ページの全部の年号を機械学習 SVM の入力とするため，それぞれの年号の順番が重要な特徴であると考える．これを利用し，法則ページの全部の年号を出現した順に，番号を付与する．例えば,「1871 年頃には着想を得ていたとされ、1923 年に文章化、完全な定式化は弟子の [[ピグー]] によって公刊された。」という文だと，年号「1871 年」を「1 番」,「1923 年」を「2 番」の出現順に番号を付与する．

3.2.2 法則対の抽出

Wikipedia の法則ページから変遷の関係にある法則対を抽出する．法則ページから抽出した基本法則と関連法則 (3.1.1 節) の対は変遷の関係にある法則対が多い．そのため，法則対の抽出では基本法則と関連法則の対から変遷の関係にある法則対を抽出する．法則対の抽出の例を図 3.5 に示す．本研究では，変遷の関係にある法則対をルーツ法則と派生法則と呼ぶ．このとき，法則年号の早い方はルーツ法則になる．これを表 1.1 の例から説明すると，法則対『「決定理論 (1670)」「ゲーム理論 (1928)」』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「決定理論」の発見年 (1670 年) が法則「ゲーム理論」の発見年 (1928 年) より早いため，法則「決定理論」はルーツ法則になる．

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 つの手法を以下に示す．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連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連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．

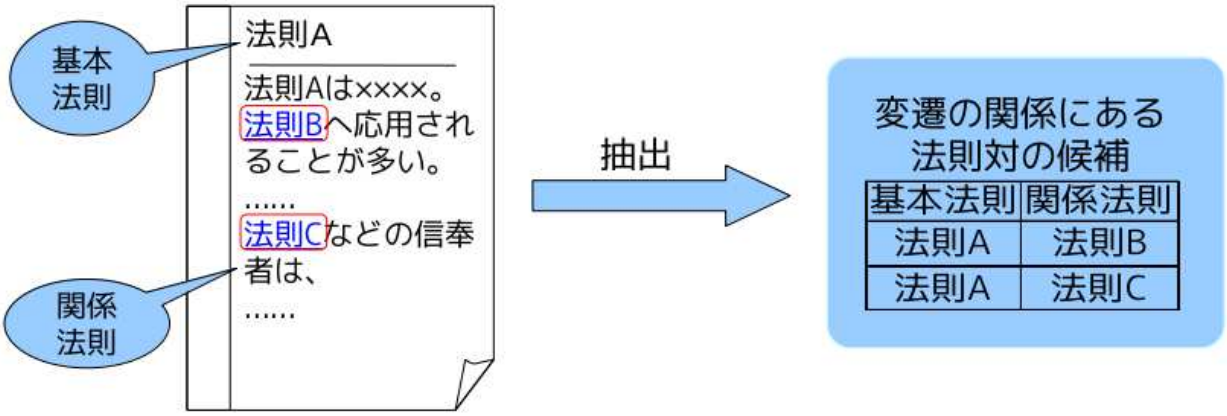

図 3.5: 法則対の抽出の例

素性の設定

ここで，手法 B2 で利用する素性を表 3.4 に示す．

表 3.4: 手法 B2 で利用した素性

素性 ID	内容
f4	法則対の名前類似度
f5	法則対は双方向法則対であるかどうか

法則対の抽出で用いる素性は,「法則対の名前類似度」,「法則対は双方向法則対であるかどうか」である．以下で，この 2 つの素性の設定について説明する．

法則対の名前類似度 法則対の法則名の類似度を計算する．変遷の関係にある法則対は法則名の語尾の一文字が一致している場合が多い．この特徴を利用し，法則名の語尾の一文字を比較することで，法則名の類似度を算出する．法則名の語尾が一致する場合に，この法則対を類似度の高いものとし，法則名の語尾が一致しない場合に，法則対を類似度の低いものとする．例えば，法則対「決定理論，ゲーム理論」の場合では，法則「決定理論」と「ゲーム理論」の語尾が一致しているため，この素性に対応する素性ベクトルの次元の値を 1 にセットする．逆に，法則対「擬似乱数,

制御理論」の場合では，法則「疑似乱数」と「制御理論」の語尾が一致していないため，この素性に対応する素性ベクトルの次元の値を 0 にセットする．

法則対は双方向法則対であるかどうか 法則対は双方向法則対であるかどうかを判断する．双方向法則対とは，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 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 3.5 に示す．法則対は双方向法則対である場合，変遷の関係にある可能性が高い．この特徴を利用し，法則対が双方向法則対であるかどうかを判断することで，変遷の関係を決める．例えば，法則対『「決定理論」「ゲーム理論」』の場合では，法則「決定理論」と「ゲーム理論」が双方向法則対であるため，この素性に対応する素性ベクトルの次元の値を 1 にセットする．そうでない場合は，この素性に対応する素性ベクトルの次元の値を 0 にセットする．

表 3.5: 双方向法則対の例

基本法則	関連法則
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

3.2.3 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．これを図 3.6 の例から説明すると，法則年号の抽出処理で取り出した法則の発見年の情報 (例えば,「法則 A(1680 年)」,「法則 B(1930 年)」) を，法則対の抽出処理で取り出した変遷の関係にある法則対の情報 (例えば,『「法則 A」「法則 B」』) と合成することで，法則の変遷情報 (『「法則 A(1680 年)」「法則 B(1930 年)」』) 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．

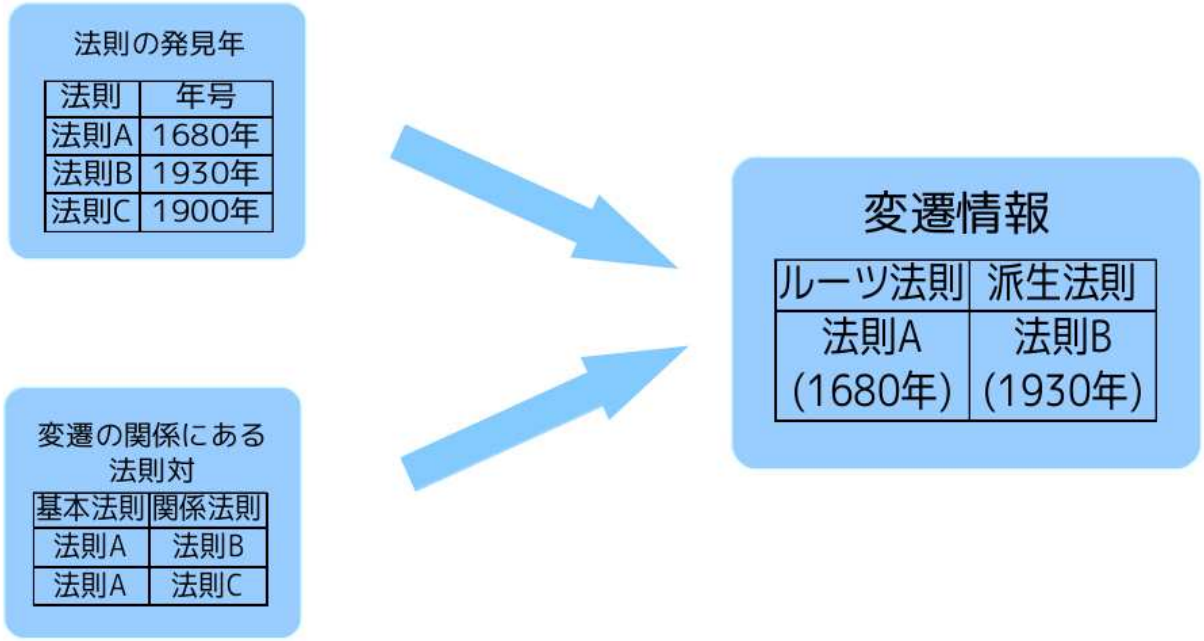

図 3.6: 変遷情報の抽出の例

3.3 サポートベクターマシン

本研究では，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)．ここで，村田 [8] の手法を参考に，サポートベクトルマシン法について説明する．サポートベクトルマシン法は，空間を超平面で分割することにより 2 つの分類からなるデータを分類する手法である．このとき，2 つの分類が正例と負例からなるものとすると，学習データにおける正例と負

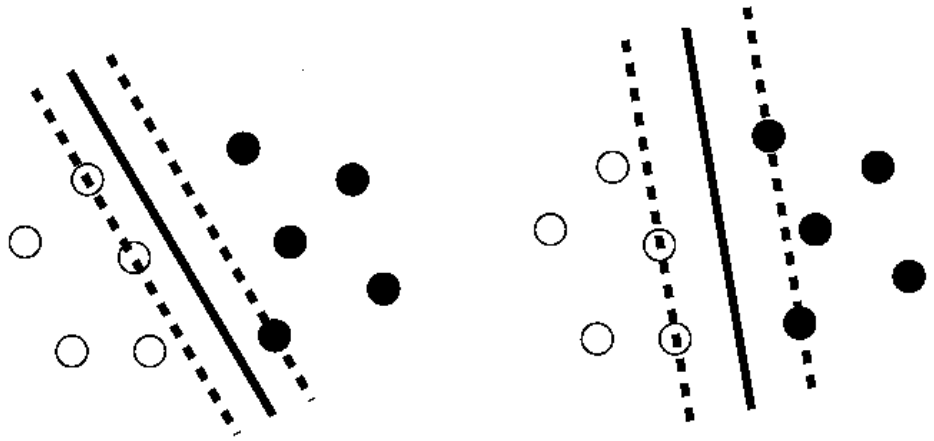

図 3.7: マージン最大化

例の間隔 (マージン) が大きいもの (図 3.7 参照[1]) ほどオープンデータで誤った分類をする可能性が低いと考えられ，このマージンを最大にする超平面を求めそれを用いて分類を行なう．基本的には上記のとおりであるが，通常，学習データにおいてマージンの内抽出の手順部領域に少数の事例が含まれてもよいとする手法の拡張や，超平面の線形の部分を非線型にする拡張 (カーネル関数の導入) がなされたものが用いられる．この拡張された方法は，以下の識別関数を用いて分類することと等価であり，その識別関数の出力値が正か負かによって二つの分類を判別することができる．

$$
\begin{aligned}
f(\mathbf{x}) &= sgn\left(\sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \\
b &= -\frac{max_{i,y_i=-1} b_i + min_{i,y_i=1} b_i}{2} \\
b_i &= \sum_{j=1}^{l} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i)
\end{aligned}
\tag{3.1}
$$

ただし，$\mathbf{x}$ は識別したい事例の文脈 (素性の集合) を，$\mathbf{x}_i$ と $y_i (i = 1, ..., l, y_i \in \{1, -1\})$ は学習データの文脈と分類先を意味し，関数 sgn は，

[1]図の白丸，黒丸は，正例，負例を意味し，実線は空間を分割する超平面を意味し，破線はマージン領域の境界を表す面を意味する．

$$sgn(x) = \begin{array}{ll} 1 & (x \geq 0) \\ -1 & (otherwise) \end{array} \tag{3.2}$$

であり，また，各 α_i は式 (3.4) と式 (3.5) の制約のもと式 (3.3) の $L(\alpha)$ を最大にする場合のものである．

$$L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}) \tag{3.3}$$

$$0 \leq \alpha_i \leq C \ (i = 1, ..., l) \tag{3.4}$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{3.5}$$

また，関数 K はカーネル関数と呼ばれ，様々なものが用いられるが本論文では以下の多項式のものを用いる．

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^d \tag{3.6}$$

C, d は実験的に設定される定数である．本論文ではすべての実験を通して C を 1 に d を 2 に固定した．ここで，$\alpha_i > 0$ となる $\mathbf{x}_i$ は，サポートベクトルと呼ばれ，通常，式 (3.1) の和をとっている部分はこの事例のみを用いて計算される．つまり，実際の解析には学習データのうちサポートベクトルと呼ばれる事例のみしか用いられない．

第4章　実験

本章では，まず前処理，実験データについて説明する．また，法則年号，法則対，法則の変遷情報の抽出性能および抽出した例を示す．最後に，実験の結果に対し考察を行う．なお，前処理において法則名の取り出し，実験データの作成は人手で行い，それ以降の変遷情報の抽出は自動で行う．変遷情報の自動抽出手法に関する詳細な説明は第3章にある．

4.1　前処理

実験には，Wikipediaから2010年5月26日にダウンロードしたWikipediaの日本語ページを利用する．そのデータに対し以下の手順で前処理を行う．

手順1 ダウンロードしたページに対し，法則ページの抽出を行う．

手順2 抽出した法則ページのデータから，法則ページごとのタイトル名を抽出し，法則名リストを作成する．

手順3 抽出した法則ページのデータから，手順2で作成した法則名リストを用いて，法則ページにある基本法則と関連法則の対を抽出し，法則対リストを作成する．

法則ページの抽出

法則の変遷情報の抽出を行うために，前処理としてWikipediaから法則を記載したページを抽出する必要がある．そのため，Wikipediaから法則ページの抽出を行う．本研究で取り扱うWikipediaのページは，図4.1に示しているようなXMLファイルのデータ構造となっている．そのうち，変遷情報の抽出で利用するXMLタグの説明を表4.1にまとめる．

法則ページの抽出では，図4.1のようなXMLファイルを入力とし，<title>と</title>で囲まれたページのタイトル名は，語尾が[分布][分類][泳動][原理][現象][数][効果] [公理][理

<title>決定理論</title>
　　<id>678</id>
<revision>
　　<id>30526938</id>
　　<timestamp>2010-02-14T05:05:43Z</timestamp>
　　<contributor><ip>121.103.206.248</ip></contributor>
　　<text xml:space="preserve">'''決定理論'''（けっていりろん、{{lang-en-short|Decision theory}}）は、個別の[[意思決定]]について[[価値]]、[[不確かさ]]といった事柄を[[数学]]的かつ[[統計学|統計的]]に確定し、それによって「最善の意思決定」を導き出す理論。'''意思決定理論'''とも。[[ゲーム理論]]へ応用されることが多い。

== 概要 ==
決定理論の大部分は規範的である。すなわち、
最良の意思決定を特定することが目的であるため、
十分な情報を持つ理想的な意思決定者を仮定し、完全な正確さで計算し、完全に合理的に意思決定するとみなす。このような規範的手法を現実の人間の意思決定に具体的に応用することを[[決定分析]](decision analysis) と呼び、人々のよりよい意思決定を支援するツール、技法、ソフトウェアの研究などを含んでいる。この考え方から生まれた最も体系的かつ総括的なソフトウェアツールを[[意思決定支援システム]]と呼ぶ。人々が最適な振る舞いをしないことは明らかなので、それに関連して、人々が実際にはどのように意思決定するかを説明しようとする研究分野もある。規範的かつ理想的な意思決定では、実際の振る舞いを評価するための仮説を生成する。これによって2つの研究分野が密接に連携する。さらに、情報の完全性や合理性などを様々な方法で緩和した場合に、どのような意思決定がなされるかを研究したり、現実になされた意思決定を評価するといった研究もある。
・・・・・・・
　　</text>
</revision>

図 4.1: Wikipedia のページ構造

表 4.1: 手法 A2 の素性

タグ	説明
[[PPP]]	PPP というページへのリンク情報
<title>	ページのタイトル名の始まり
</title>	ページのタイトル名の終わり
<text>	ページの内容の始まり
</text>	ページの内容の終わり

論][収差][定理][転位][予想][法][価][律][線][説][式][則] と一致した場合に，そのページを法則ページとして抽出する．

4.2 実験データ

前処理で得たデータの内訳は以下のとおりである．抽出した法則ページの数は 5,061 個である．そのうち年号のある法則ページの数は合計 1,634 個である．年号のある法則ページにより作成した法則名リストには 1,634 個の法則名を収録した．年号のある法則ページから抽出した法則対は，延べで 2,074 個，異なりで 1,621 個である．

また，法則名リスト，法則対リストの例をそれぞれ表 4.2，表 4.3 に示す．

データ A とデータ B

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(異なる法則の数 : 133 個) とする．データ 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(異なる法則の数 : 137 個) とする．

4.3 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．

$$\text{再現率} = \frac{\text{手法の出力のうちの正解数}}{\text{実際の正解の数}} \tag{4.1}$$

$$\text{適合率} = \frac{\text{手法の出力のうちの正解数}}{\text{手法による出力の数}} \tag{4.2}$$

表 4.2: 法則名リスト

法則名
ジムロート転位
ブルック転位
司法的執行の理論
13 星座説
14 角数
16 進数
16 進法
2-アダマンタノンの構造式
2003 の等ラウドネス曲線
2007 年の大学別特許公開件数
2012 年人類滅亡説
2012 年地球滅亡説
栄養的分類
永久法
永字八法
永田の定理
永田の埋め込み定理
永田農法
永田法
泳法
衛星現象
衛生仮説
衛生法
液化ガス蒸気圧曲線
液気圧式
液晶式
液晶表示式

表 4.3: 法則対リスト

基本法則	関連法則
東証株価指数	浮動株基準株価指数
特殊相対性理論	一般相対性理論
特殊相対性理論	マクスウェルの方程式
特殊相対性理論	光電効果
特殊相対性理論	コンプトン効果
特殊相対性理論	ドップラー効果
特殊相対性理論	尺貫法
グラフ理論	四色定理
グラフ理論	ワーシャル-フロイド法
グラフ理論	スモール・ワールド現象
生成文法	格文法
生成文法	一般化句構造文法
生成文法	主辞駆動句構造文法
生成文法	最適性理論
カー・パリネロ法	分子動力学法
ブラック-ショールズ方程式	確率微分方程式
量子化学的手法	分子軌道法
DFPT 法	BCS 理論
共形場理論	超弦理論
シュレーディンガー方程式	ディラック方程式
計算複雑性理論	計算理論
計算複雑性理論	計算可能関数

$$F\text{ 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \tag{4.3}$$

4.4 評価

第 3 章で提案した変遷情報の自動抽出手法の性能をここで評価する．評価は法則年号の抽出，法則対の抽出，変遷情報の抽出ごとに行う．

4.4.1 法則年号の抽出性能

第 3 章の 3.2.1 節で提案した法則年号の抽出手法の性能を評価する．教師あり機械学習の実験 (手法 A2 と手法 A3 の場合) では，データ A の 133 個の法則を学習データとし，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 4.4 に示す．また，抽出した法則年号の例を表 4.5 に示す．

表 4.4: 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(85/92)	0.62(85/137)	0.74
手法 A2	0.75(69/92)	0.86(69/ 80)	0.80
手法 A3	0.68(63/92)	0.83(63/ 76)	0.75

4.4.2 法則対の抽出性能

第 3 章の 3.2.2 節で提案した法則対の抽出手法の性能を評価する．教師あり機械学習の実験 (手法 B2 の場合) では，データ A の 100 個の法則対を学習データとし，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 4.6 に示す．また，抽出した変遷の関係にある法則対の例を表 4.7 に示す．

4.4.3 変遷情報の抽出性能

第 3 章の 3.2.3 節で提案した変遷情報の抽出手法の性能を評価する．変遷情報の抽出精度を 3 つの基準で評価する．

表 4.5: 法則年号の例

法則	年号	結果
決定理論	1670 年	正解
ゲーム理論	1928 年	正解
プロスペクト理論	1979 年	正解
ホフマン転位	1871 年	正解
クルチウス転位	1890 年	正解
班田収授法	646 年	正解
三世一身法	723 年	正解
十二表法	紀元前 451 年	正解
ホルテンシウス法	紀元前 287 年	正解
リキニウス・セクスティウス法	紀元前 367 年	正解
ハッブルの法則	1929 年	正解
ドップラー効果	1842 年	正解
測量法	1949 年	正解
水路業務法	1950 年	正解
SMILES 記法	1980 年	正解
グラフ理論	1736 年	正解
計算理論	2000 年	誤り
計算可能関数	1930 年	誤り
計算複雑性理論	2006 年	誤り
計算可能性理論	1936 年	誤り
モルワイデ図法	2005 年	誤り
階乗素数	2007 年	誤り
素数	2008 年	誤り
超光速航法	2001 年	誤り
ロボット工学三原則	2058 年	誤り
バンドワゴン効果	1996 年	誤り
微分方程式	2005 年	誤り
解析関数	1938 年	誤り
ベンフォードの法則	1938 年	誤り
数学定数	1950 年	誤り

表 4.6: 法則対の抽出の結果

手法	再現率	適合率	F 値
手法 B1	1.00(60/60)	0.60(60/100)	0.75
手法 B2	0.87(52/60)	0.88(52/ 59)	0.87

表 4.7: 変遷の関係にある法則対の例

ルーツ法則	派生法則	結果
決定理論	ゲーム理論	正解
決定理論	プロスペクト理論	正解
ホフマン転位	クルチウス転位	正解
班田収授法	三世一身法	正解
十二表法	ホルテンシウス法	正解
十二表法	リキニウス・セクスティウス法	正解
超弦理論	共形場理論	正解
一般相対性理論	M 理論	正解
揺動散逸定理	相反定理	正解
制御理論	H ∞制御理論	正解
擬似乱数	制御理論	誤り
実数	はさみうちの原理	誤り
フェーン現象	日本架空説	誤り
自然数	公理	誤り
超光速航法	ロボット工学三原則	誤り
地球空洞説	オイラーの公式	誤り
陳景潤の定理	素数	誤り
偶数	幸運数	誤り
干渉法	サニャック効果	誤り
数学的帰納法	ガウス整数	誤り

基準 1 抽出した法則対が正しい変遷の関係であり，かつ，抽出した法則年号も正しい場合，正解と判断する基準.

基準 2 抽出した法則対が正しい変遷の関係である場合，正解と判断する基準．すなわち，抽出した法則年号が間違っていても，良いとする.

基準 3 抽出した法則対が変遷の関係である場合，正解と判断する基準．すなわち，抽出した法則年号によるルーツと派生の順番は判定しない.

変遷情報の抽出性能を表 4.8 に示す．また，変遷情報の例を表 4.9 に示す.

表 4.8: 変遷情報の抽出の結果

手法	基準	再現率	適合率	F 値
C1 (A1 + B1)	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 (A2 + B1)	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 (A3 + B1)	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 (A1 + B2)	1	0.77(24/31)	0.41(24/ 59)	0.54
	2	0.77(24/31)	0.41(24/ 59)	0.54
	3	0.87(52/60)	0.88(52/ 59)	0.87
C5 (A2 + B2)	1	0.55(17/31)	0.71(17/ 24)	0.62
	2	0.55(17/31)	0.71(17/ 24)	0.62
	3	0.87(52/60)	0.88(52/ 59)	0.87
C6 (A3 + B2)	1	0.39(12/31)	0.71(12/ 17)	0.50
	2	0.39(12/31)	0.71(12/ 17)	0.50
	3	0.87(52/60)	0.88(52/ 59)	0.87

4.5 考察

表 4.4 と表 4.6 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 値を得ていることがわかる．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.

表 4.9: 変遷情報の例

ルーツ法則	派生法則
決定理論 (1670 年)	ゲーム理論 (1928 年)
決定理論 (1670 年)	プロスペクト理論 (1979 年)
ホフマン転位 (1871 年)	クルチウス転位 (1890 年)
班田収授法 (646 年)	三世一身法 (723 年)
十二表法 (紀元前 451 年)	ホルテンシウス法 (紀元前 287 年)
超弦理論 (1960 年)	共形場理論 (1980 年)
シュレーディンガー方程式 (1926 年)	ディラック方程式 (1928 年)
ゼーベック効果 (1821 年)	相反定理 (1931 年)
六法 (1874 年)	景観緑三法 (2004 年)
制御理論 (1950 年)	H ∞制御理論 (1980 年)
ヤード・ポンド法 (1824 年)	度量衡法 (1891 年)
ハッブルの法則 (1929 年)	ドップラー効果 (1842 年)
原子価殻電子対反発則 (1939 年)	オクテット則 (1916 年)
ホフマン転位 (1871 年)	クルチウス転位 (1890 年)
ZND 理論 (1940 年)	CJ 理論 (1890 年)
一般化句構造文法 (1970 年)	語彙機能文法 (1970 年)

次に表 4.8 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，基本法則と関連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，この性能が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．

第5章　素性分析

法則の変遷情報の抽出で利用した素性が提案手法の性能向上に有効であるか否かを調査するため，法則年号の抽出，法則対の抽出ごとに利用した素性に対する分析を行う．素性分析では全ての素性を一つずつ省いて実験し，ある素性を省いたときの性能と全部の素性を利用するときの性能を比較することで，素性の有効性を調査する．実験は 10 分割クロスバリデーションを用いて行う．具体的には，クロスバリデーションの結果により，ある素性に対し，もしその素性を省いたときの性能が全部の素性を利用するときの性能より低い場合，その素性が有効であると判断し，もしその素性を省いたときの性能が全部の素性を利用するときの性能より高い場合，その素性が有効でないと判断する．ここで，性能評価に用いる基準を F 値とする．

5.1　法則年号の抽出

法則年号の抽出において教師あり機械学習に基づく手法は A2，A3 である．ここで，手法 A2 と A3 に対し素性分析を行い，その結果を表 5.2 と表 5.3 に示す．素性の説明について，表 5.1 に示す．

表 5.1: 素性の説明

素性 ID	説明
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

手法 A2 では,「年号前後の文字列 (f1)」と「文頭から年号までの文の長さ (f2)」の 2 つの素性を用いた．表 5.2 の結果により，素性「年号前後の文字列 (f1)」は省いたときの F 値 (0.74) が全部の素性を利用するときの F 値 (0.71) より性能が上がることから，この素性は提案手法の性能向上に有効でないことがわかる．一方，素性「文頭から年号までの文の長さ (f2)」の場合は，省いたときの F 値 (0.69) が全部の素性を利用するときの F 値

表 5.2: 手法 A2 の素性分析の結果

利用した素性	省いた素性	再現率	適合率	F 値
f1	f2	0.63	0.77	0.69
f2	f1	0.79	0.71	0.74
全素性	—	0.65	0.78	0.71

表 5.3: 手法 A3 の素性分析の結果

利用した素性	省いた素性	再現率	適合率	F 値
f1,f2	f3	0.43	0.75	0.54
f1,f3	f2	0.63	0.77	0.70
f2,f3	f1	0.79	0.71	0.74
全素性	—	0.70	0.81	0.75

(0.71) より性能が下がることから，この素性は提案手法の性能向上に有効であることがわかる．

手法 A3 では,「年号前後の文字列 (f1)」と「文頭から年号までの文の長さ (f2)」と「年号の順番 (f3)」の 3 つの素性を用いた．表 5.3 の結果により，素性「年号前後の文字列 (f1)」，素性「文頭から年号までの文の長さ (f2)」，素性「年号の順番 (f3)」の 3 つの素性はそれぞれ省いたときの F 値 (0.74，0.70，0.54) が全部の素性を利用するときの F 値 (0.75) よりも性能が下がることから，この 3 つの素性は全て提案手法の性能向上に有効であることがわかる．特に，素性「年号の順番 (f3)」を省いたときの性能の差が最も大きく，素性「年号の順番 (f3)」は最も効果があるといえる．

5.2 法則対の抽出

法則対の抽出において教師あり機械学習に基づく手法は B2 である．ここで，手法 B2 に対し素性分析を行い，その結果を表 5.5 に示す．素性の説明について，表 5.4 に示す．

表 5.4: 素性の説明

素性 ID	説明
f4	法則対の名前類似度
f5	法則対は双方向法則対であるかどうか

表 5.5: 手法 B2 の素性分析の結果

利用した素性	省いた素性	再現率	適合率	F 値
f4	f5	0.75	0.85	0.80
f5	f4	0.92	0.74	0.82
全素性	—	0.92	0.74	0.82

手法 B2 では,「法則対の名前類似度 (f4)」と「法則対は双方向法則対であるかどうか (f5)」の 2 つの素性を用いた．表 5.5 の結果により，素性「法則対の名前類似度 (f4)」は省いたときの F 値 (0.82) が全部の素性を利用するときの F 値 (0.82) との性能が相等することから，この素性は提案手法の性能向上に有効であるとはいえない．一方，素性「法則対は双方向法則対であるかどうか (f5)」の場合は，省いたときの F 値 (0.80) が全部の素性を利用するときの F 値 (0.82) より性能が下がることから，この素性は提案手法の性能向上に有効であることがわかる．

第6章　Wikipediaにおける法則以外の変遷に対する調査

本研究では Wikipedia からの法則の変遷情報の抽出を主に行ったが，法則だけに限らず法則以外のもの (例えば：人，文化) の変遷情報の抽出が考えられる．また，法則の変遷情報の抽出で提案した手法が，法則以外のものの変遷情報の抽出に応用できるか否かという提案手法の汎用性を検証したい．このことから，Wikipedia からの法則以外のものに対して変遷関係の調査を行う．本章では，調査の手順，調査の結果について述べる．

6.1　調査の手順

調査の手順は以下のとおりである．

手順 1 実験データとして Wikipedia から 2012 年 10 月 27 日時点の全日本語ページ (9GB) をダウンロードする．

手順 2 ダウンロードしたページに対し，その内訳 (年号を含むページの数，年号が最初の段落に出現したページの数) を調査する．

手順 3 ダウンロードした Wikipedia のページをカテゴリごとに分類し，各カテゴリに属するページの数を求め，カテゴリの一覧表を作成する．

手順 4 一覧表の上位にある代表的なカテゴリを選び，法則の変遷情報の抽出で提案したヒューリスティックルールに基づく手法を用いて，機械でカテゴリごとの変遷情報を抽出する．

手順 5 抽出したカテゴリごとの変遷情報をランダムに 10 件ずつ取りだし，人手評価を行う．

6.2 調査の結果

ダウンロードしたWikipediaのページに対して，年号を含むページの数，年号が最初の段落に出現したページの数を調査した．その結果を表6.1に示す．

表 6.1: Wikipedia ページの内訳

項目	総数
全日本語ページ	2,288,325
年号を含むページ	1,364,576
年号が最初の段落に出現したページ	245,331

年号が最初の段落に出現したページ (総数:245,331) をカテゴリごとに分類した結果，総数 76,774 個のカテゴリを得た．そのカテゴリの一覧を表6.2に示す．表6.2により，Wikipediaには人に関する記事が最も多いことがわかった．例えば，1位にある「存命人物」に属するページの数は他のカテゴリより比較的に多く，37,435 件に達すること．

表6.2のカテゴリから,「存命人物」,「日本の映画作品」,「アニメソング」を代表的なカテゴリとして選び，機械でカテゴリごとの変遷情報を抽出する．抽出したカテゴリごとの変遷情報の例を表6.3，表6.4，表6.5に示す．

また，カテゴリ「存命人物」に対して，機械で抽出した変遷情報から，ランダムに10件取りだし，人手評価を行った．カテゴリ「存命人物」の人手評価の例を表6.6，人手評価の結果を表6.7に示す．

6.3 まとめ

Wikipedia からの法則以外の記事に対して変遷関係の調査を行った．データとしてWikipedia の 2012 年 10 月 27 日時点の全日本語ページを利用した．総数 245,331 個あるページをカテゴリごとに分類し，総数 76,774 個のカテゴリを得た．法則の変遷情報の抽出で提案したヒューリスティックルールに基づく手法を用いて，Wikipediaから，カテゴリ「存命人物」,「日本の映画作品」,「アニメソング」の変遷情報を抽出した．また，カテゴリ「存命人物」に対して人手評価を行った．人手評価の結果，年号の正解率が 1.0，変遷関係にある対の正解率が 0.70，年号と変遷関係を両方抽出する場合の正解率は0.70となった．この結果から，法則の変遷情報の抽出で提案した手法は法則以外のものの変遷情報の抽出に応用可能であることがわかった．

表 6.2: カテゴリの一覧表

順位	カテゴリ	ページ数
1	存命人物	37,435
2	学校記事	3,133
3	アメリカ合衆国の映画作品	2,578
4	アメリカ合衆国の俳優	2,215
5	日本の映画作品	2,047
6	東京都出身の人物	2,003
7	サラブレッド	1,526
8	アメリカ合衆国の野球選手	1,361
9	ドラマ映画	1,224
10	ベスト・アルバム	1,131
11	日本の俳優	1,126
12	アニメソング	1,118
13	1980 年生	1,061
14	1981 年生	1,048
15	1982 年生	1,023
16	文学を原作とする映画作品	1,019
17	カリフォルニア州の人物	1,019
18	ニューヨーク市出身の人物	1,012
19	1983 年生	1,006
20	美少女ゲーム	996
21	1979 年生	992
22	1984 年生	974
23	長大な音楽作品名	955
24	1985 年生	930
25	1978 年生	929
…	…	…

表 6.3: 存命人物の例

人物 A	人物 B
西科仁 (1978 年)	ぢゃいこ (1981 年)
西科仁 (1978 年)	有賀明美 (1977 年)
窪塚俊介 (1981 年)	板垣恭一 (1964 年)
三輪ひとみ (1978 年)	三輪恵未 (1987 年)
三輪ひとみ (1978 年)	佐々木浩久 (1961 年)

表 6.4: 日本の映画の例

映画 A	映画 B
幕末太陽傳 (1957 年)	太陽の季節 (1955 年)
幕末太陽傳 (1957 年)	田園に死す (1974 年)
幕末太陽傳 (1957 年)	蜘蛛巣城 (1957 年)
幕末太陽傳 (1957 年)	少年探偵団 (1975 年)
転々(2007 年)	図鑑に載ってない虫 (2007 年)

表 6.5: アニメソングの例

アニメソング A	アニメソング B
Check my soul(2012 年)	真夏のフォトグラフ (2011 年)
Check my soul(2012 年)	リスアニ!(2010 年)
流れ星☆ (2003 年)	大切な願い (2003 年)
青空 loop(2007 年)	空中迷路 (2007 年)
世界の果てに君がいても (2011 年)	世界で一番恋してる (2011 年)

表 6.6: 「存命人物」の人手評価の例

	人物 A	人物 B	人物 A の年号	人物 B の年号	変遷の関係	年号と関係
1	田口節子 (1981 年)	丸岡正典 (1979 年)	正解	正解	正解	正解
2	蓑輪単志 (1959 年)	大友康平 (1956 年)	正解	正解	正解	正解
3	松田慶三 (1975 年)	平野勝美 (1965 年)	正解	正解	誤り	誤り
4	砂子塾長 (1964 年)	砂子義一 (1932 年)	正解	正解	正解	正解
5	志村有弘 (1941 年)	松本寧至 (1931 年)	正解	正解	正解	正解
6	清水美那 (1982 年)	大嶋拓 (1963 年)	正解	正解	誤り	誤り
7	水江慎一郎 (1965 年)	泉見洋平 (1972 年)	正解	正解	正解	正解
8	峰一也 (1950 年)	黒田透 (1955 年)	正解	正解	正解	正解
9	久野秀隆 (1987 年)	横上拓哉 (1986 年)	正解	正解	誤り	誤り
10	吉川真司 (1960 年)	栄原永遠男 (1946 年)	正解	正解	正解	正解

表 6.7: 「存命人物」の人手評価の結果

カテゴリ	年号の正解率	変遷関係の正解率	年号と関係の正解率
存命人物	1.00(20/20)	0.70(7/10)	0.70(7/10)

第7章　今後の課題

本研究ではWikipediaから法則の変遷情報の抽出を行ったが，またいくつかの問題が残っている．本章では，残っている問題を今後の課題として以下にまとめる．

- 本研究では情報抽出の性能を評価する基準として，再現率と適合率の調和平均であるF値を用いたが，F値で評価をした方がよいのか，または適合率で評価をした方がよいのかという点について，さらに検討したい．
- 取得した変遷情報をより効果的に表示するための可視化ツールを開発したい．
- 今回の実験データは人手で判定したが，実験の客観性を保つため，より複数の判定者による判定を行いたい．
- 素性分析の結果により，手法A2の素性f1が提案手法の性能向上に有効でないことがわかった．その素性を省くことで変遷情報の抽出の性能がよくなるか否かを検証したい．
- 法則以外のものに対する実験が少ない．法則以外のものに対する実験を増やしたい．
- 実用化するため，変遷情報の自動抽出機能をパッケージ化したい．

第8章　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変遷情報の抽出を行った．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，基本法則と関連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．ヒューリスティックルールに加え教師あり機械学習を利用する手法で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．

また，提案手法の汎用性を検証すべく，提案手法を用いて法則以外のものの変遷情報の抽出を試みた．その結果，提案手法は法則以外のものの変遷情報の抽出に応用可能であることがわかった．

最後に，情報抽出の評価基準，可視化ツールの作成，実験データの人手判定，素性の実験，法則以外の実験，パッケージ化を今後の課題として取り上げたいと考える．

謝辞

本研究を進めるに当たり，終始に渡り研究の進め方や本論文の書き方など，細部にわたる御指導を頂きました，鳥取大学工学部知能情報工学科計算機工学講座 C の村田真樹教授に心から御礼申し上げます．また，本研究を進めるに当たり，御指導，御助言を頂きました，村上仁一准教授，徳久雅人講師に心から御礼申し上げます．加えて，種々の御助言を龍谷大学理工学部数理情報学科の馬青教授に頂きました．ここに深く感謝いたします．その他様々な場面で御助言を頂いた計算機工学講座 C 研究室の皆様に感謝の意を表します．

参考文献

[1] Sanako Hori, Masaki Murata, Masato Tokuhisa and Qing Ma: “Automatic Extraction of Historical Transition in Researchers and Research Topics”, In Processing of the 7th IEEE Conference on Natural Language Processing and Knowledge Engineering (NLPKE 2011), pp.296-299, 2011.

[2] 堀さな子, 村田真樹, 徳久雅人, 馬青: “パターンと機械学習を用いた大規模テキストからの変遷情報の抽出と分類”, 言語処理学会第 19 回年次大会発表論文集, pp.592–595, 2013.

[3] 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.

[4] 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.

[5] 山田一郎, 呉鍾勳, 鳥澤健太郎, 黒田航, 風間淳一, 村田真樹: “Wikipedia を利用した日本語 WordNet への用語追加の検討”, 言語処理学会第 16 回年次大会発表論文集, pp.948–951, 2010.

[6] 戸田智子, 福田直樹, 石川博: “Blog 記事のクラスタリングに基づいたカテゴリ別話題変遷パタンの抽出”, 電子情報通信学会第 18 回データ工学ワークショップ, DEWS2007, A8-Blog, 2007.

[7] 野田陽平, 清田陽司, 中川裕志: “意外性のある知識発見のための Wikipedia カテゴリ間の関係分析”, 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-A803-02, 2009.

[8] 村田真樹: “機械学習手法を用いた日本語格解析—教師信号借用型と非借用型，さらには併用型—”, 情報処理学会研究報告, 自然言語処理研究報告 2001(69), pp.113–120, 2001.

[9] Wikipedia: http://ja.wikipedia.org/wiki/

[10] TinySVM: http://chasen.org/ taku/software/TinySVM/

[11] Liangliang Fan, Masaki Murata, Masato Tokuhisa and Qing Ma: “Extraction of Historical Transition in Legal and Scientific Laws from Wikipedia”, In Proceedings of the 8th Conference on Natural Language Processing and Knowledge Engineering (NLPKE 2012), pp.144-155, 2012.

付録 1

- 修士論文予稿 (2013 年 8 月)
- International Journal of Advanced Intelligence(2013 年 7 月，発行予定)
- 言語処理学会第 18 回年次大会発表論文集 (2012 年 3 月)