# 概要

オンライン百科事典である Wikipedia は，近年様々な研究に利用されている．しかし，Wikipedia の利用者に対し支援を行っている研究は少ない．Wikipedia は誰もが編集可能であるという特徴を持ち，大量の情報を得ることに成功しているが，それにともない，その中から目的の情報を探すための負担は，情報が増えることに比例して大きくなっていくため，効率の良い情報収集のための支援が求められている．

本研究では，セクション名に注目し，セクションに特定の情報が存在するか否かを表す情報タグの付与を行うことによって，Wikipedia 利用者の支援を行った．情報タグは，教師あり機械学習である SVM や，パターンマッチングによって付与を行う．また，先行研究を参考に，Wikipedia から機械学習で使用する教師データを自動的に生成する，推定教師データと呼ばれる教師データを生成し，その推定教師データを利用した教師学習による情報タグ付与の評価を行った．

その結果，推定教師データを利用した SVM は，F 値が 0.524（再現率 0.806，適合率 0.492）であり，手作業で教師データを作成した SVM の F 値 0.612（再現率 0.511，適合率 0.536）よりも低いが，パターンマッチングの F 値 0.554（再現率 0.937，適合率 0.3941）とほぼ同程度であることがわかった．ここで，既存のセクションから情報の有無を判断する，セクション名利用による方法の F 値が極端に低い 0.129（再現率 0.069，適合率 0.818）であることから，既存のセクション名では情報が不足しており，セクション名に何らかの支援が必要であることが明らかになった．また，推定教師データと教師データとの組み合わせた SVM，推定教師データを利用した Stacking，推定教師データ数，教師データ数の変化による F 値の調査を行うことによって推定教師データ利用の可能性を調査した．

# 目 次

**第 1 章　はじめに　1**

**第 2 章　関連研究　2**

2.1 Wikipedia を用いた用語説明のモデル化と事典的検索への応用 . . . . . 2
2.2 情報抽出 . . . . . 2
2.2.1 Wikipedia からの大規模な上位下位関係の獲得 . . . . . 2
2.2.2 Wikipedia カテゴリネットワークからの意外性のある関係性の抽出 3
2.3 Wikipedia に対する支援 . . . . . 3
2.3.1 Wikipedia におけるミッシングリンクの自動発見手法 . . . . . 3
2.3.2 Wikipedia の編集履歴に基づく記事の信頼性導出 . . . . . 3
2.4 分析 . . . . . 3
2.4.1 Wikipedia における編集者の活動分析 . . . . . 3
2.4.2 ウィキペディア記事閲覧回数の特徴分析 . . . . . 4
2.5 利用可能な Wikipedia の情報 . . . . . 4

**第 3 章　Wikipedia への情報付与　5**

3.1 問題設定 . . . . . 5
3.2 実験環境 . . . . . 5
3.3 Support Vector Machine への適用 . . . . . 5
3.3.1 Support Vector Machine . . . . . 5
3.3.2 素性 . . . . . 7
3.4 法則ページ抽出方法 . . . . . 8
3.4.1 Wikipedia の構成 . . . . . 8
3.4.2 法則名抽出 . . . . . 8
3.4.3 法則内容抽出 . . . . . 9
3.4.4 法則内容の分割 . . . . . 9
3.4.5 情報タグ候補の作成方法 . . . . . 10

3.5 情報タグ付与方法 . . . . . 10
3.5.1 手作業で教師データを利用したSVM . . . . . 10
3.5.2 推定教師データを利用したSVM . . . . . 10
3.5.3 パターンマッチング . . . . . 11
3.5.4 セクション名利用 . . . . . 11
3.5.5 全てに情報タグを付与 . . . . . 11
3.6 教師データ作成手法 . . . . . 11
3.6.1 手作業による教師データの作成 . . . . . 11
3.6.2 推定教師データの作成 . . . . . 11

**第4章 実験** **13**
4.1 事前準備 . . . . . 13
4.1.1 法則名抽出 . . . . . 13
4.1.2 法則内容抽出 . . . . . 14
4.1.3 法則内容分割 . . . . . 14
4.2 情報タグ候補の決定 . . . . . 15
4.3 教師データ作成 . . . . . 16
4.4 セクション名の付与結果 . . . . . 17

**第5章 評価** **18**
5.1 評価値の計算 . . . . . 18
5.1.1 F値の計算 . . . . . 18
5.1.2 交差検定 . . . . . 18
5.1.3 ブートストラップ法 . . . . . 19
5.2 情報タグ「歴史」の評価 . . . . . 19
5.2.1 比較手法 . . . . . 19
5.2.2 比較結果 . . . . . 20
5.2.3 有意差検定 . . . . . 20
5.3 証明，例，定義の評価 . . . . . 21
5.3.1 評価値算出方法 . . . . . 21
5.3.2 評価結果 . . . . . 21
5.3.3 F値の算出 . . . . . 22
5.3.4 有意差検定 . . . . . 22

5.4 教師データと推定教師データの組み合わせ . . . . . 23
5.4.1 Stacking . . . . . 23
5.4.2 推定教師データと教師データの合成 . . . . . 23
5.5 教師データ数ごとの評価値 . . . . . 24
5.5.1 教師データ . . . . . 24
5.5.2 推定教師データ . . . . . 25

**第 6 章 考察** **26**
6.1 支援の必要性 . . . . . 26
6.1.1 教師データからの考察 . . . . . 26
6.1.2 F 値からの考察 . . . . . 29
6.1.3 情報タグ付与結果からの考察 . . . . . 30
6.2 推定教師データ . . . . . 30
6.2.1 教師データとの比較 . . . . . 30
6.2.2 パターンマッチング手法との比較 . . . . . 31
6.2.3 情報タグごとの比較 . . . . . 31
6.2.4 利点と欠点 . . . . . 32
6.3 各手法の特徴 . . . . . 32
6.4 手法の組み合わせ . . . . . 33
6.4.1 他手法と教師あり機械学習の組み合わせ . . . . . 33
6.4.2 教師データと推定教師データ組み合わせ . . . . . 33

**第 7 章 おわりに** **34**

# 図 目 次

3.1 分割前データ . . . . . 9
3.2 分割後データ . . . . . 10

4.1 付与結果例 . . . . . 17

6.1 冒頭のセクション . . . . . 28
6.2 歴史と類義語のセクション名の例 . . . . . 28
6.3 最も効果のある例 . . . . . 29
6.4 情報混在例 . . . . . 29
6.5 フラッシュ法 . . . . . 31
6.6 証明のセクション例 . . . . . 32

# 表 目 次


3.1 素性例 . . . . . 7
3.2 分類例 . . . . . 12

4.1 獲得法則名例 . . . . . 13
4.2 排除した法則名例 . . . . . 13
4.3 抽出例 . . . . . 14
4.4 抽出例 2 . . . . . 14
4.5 情報タグ候補 . . . . . 15
4.6 推定教師データ数 . . . . . 16
4.7 教師データ数 . . . . . 16

5.1 手法の比較 . . . . . 20
5.2 ブートストラップ法 . . . . . 20
5.3 評価：証明 . . . . . 21
5.4 評価：定義 . . . . . 21
5.5 評価：例 . . . . . 21
5.6 F 値の比較 . . . . . 22
5.7 再現率適合率の比較 . . . . . 22
5.8 ブートストラップ法 2 . . . . . 23
5.9 スタッキング . . . . . 23
5.10 教師＋推定教師 . . . . . 24
5.11 教師データ数ごとの評価 . . . . . 24
5.12 推定教師データ数ごとの評価 . . . . . 25

6.1 歴史の情報を含むセクションの内訳 . . . . . 26
6.2 情報タグ「:歴史」の内訳 . . . . . 27
6.3 情報タグ付与結果 . . . . . 30

6.4 法則関連ページ予測数 . . . . . 30
6.5 手法の特徴 . . . . . 33

# 第1章　はじめに

近年，Wikipedia と呼ばれるオンライン百科事典が急速な成長を見せている．誰もが編集可能であるという特徴を持った事典は，多方面に精通した辞書を得ることに成功しているが，その中から目的の情報を探すための負担は，情報が増えることに比例して大きくなっていくため，効率の良い情報収集のための支援が求められている．Wikipedia では，記事が章や節で区切られている．本研究では，その章や節のことをセクションと呼び，セクションに与えられた見出しをセクション名と呼ぶ．利用者はセクション名を頼りに情報を絞り，目的の情報を探す．しかし，セクション名は厳格な基準がなく文章作成者の判断によって決められているため，統一されたセクション名はつけられていない．また，冒頭のセクションにはセクション名がつけられていない．そのため利用者は，目的のセクション名を見つけられなかった場合，どのセクションに必要な情報が書かれているか想定できないため，そのページのセクション全てをチェックしなければならなくなる．藤井らの研究 [1] では，セクション名に注目して Wikipedia から教師データを作成し，検索した単語の要約を行っているが，他の手法との比較実験は行っていない．

そこで本研究では，利用者の支援となる情報タグをセクションに付与する方法を提案すると共に，種々の手法を利用した情報タグ付与を試み，各手法の特徴を明らかにする．情報タグとは，どのような情報が各セクションに存在するかを示すものである．Wikipedia の利用者は，情報タグによりどのような情報がどのセクションに存在しているかを容易に知ることができる．本論文の構成は以下の通りである．第 2 章ではこれまでの Wikipedia に対する研究を説明し，利用可能な Wikipedia の情報を説明する．第 3 章では本研究における，情報タグの付与方法と情報付与に利用する技術を説明する．第 4 章では，Wikipedia に情報タグの付与を行う．第 5 章では，情報タグの付与結果に対する手法ごと，分類ごとの評価を行う．また，推定教師データを利用した Stacking，教師数による F 値の変化を調査することによって，推定教師データの分析を行う．第 6 章で考察を行い推定教師データの効果的な利用法を考察する．第 7 章ではまとめを行う．

# 第2章　関連研究

Wikipediaに関する研究は多く，様々な試みがなされている．本章では，Wikipediaに対して行われてきた様々な先行研究を紹介する．また，先行研究で使用されている，利用可能なWikipediaの情報についてまとめる．

## 2.1　Wikipediaを用いた用語説明のモデル化と事典的検索への応用

第1章で紹介した藤井らの研究[1]の教師データ作成方法の説明を行う．入力された用語について検索されたテキストを入力として扱い，用語分類，観点分類を順番に実行する．用語分類は，Wikipediaの「病名」「人名」「動物名」のページ集合をページごとに学習を行い，入力されたテキストが「病名」「人名」「動物名」のうちどの事柄について記述されているかを判断する．観点分類は，例えば用語分類において「病名」に分類されたテキストを入力とする場合,「病名」に関するWikipediaのページ集合を観点（「病状」「原因」「治療」）ごとに学習を行い，そのテキストが「症状」「原因」「治療」のどの観点について記述されているかを判断する．本研究では，この二つの分類器のうち特に観点分類の技術を参考に教師データを自動的に生成する．

## 2.2　情報抽出

### 2.2.1　Wikipediaからの大規模な上位下位関係の獲得

隅田ら[2]は，WikipediaのMediaWiki構文の修飾記号を利用して獲得した上位下位関係を，Support Vector Machineによってフィルタリングを行い，高い適合率で上位下位関係を大量に獲得している．

### 2.2.2 Wikipedia カテゴリネットワークからの意外性のある関係性の抽出

野田ら [3] は，Wikipedia のカテゴリネットワークから特徴量を抽出し，機械学習を用いて意外性のある関係性の抽出を行った．しかし，少量の意外性を含む関係性を確実に抽出するのは困難であること，意外性の定義が人それぞれ違うことなど問題が残った．

## 2.3 Wikipedia に対する支援

### 2.3.1 Wikipedia におけるミッシングリンクの自動発見手法

中川ら [4] は，意味的に関係しているのにリンクが作成されていない「ミッシングリンク」と呼ばれる問題に対し，従来手法の改良を行っている．従来手法では，対象記事の関連記事内にあるリンク情報を，対象記事のリンクと比較を行い，不足しているリンクを補っていた．それに対し提案手法では，被リンク数の少ない記事に対し、経由する記事によりリンクに重み付けをして関連度を算出することで精度を向上させた．

### 2.3.2 Wikipedia の編集履歴に基づく記事の信頼性導出

鈴木ら [5] は，編集履歴を利用することによって，Wikipedia 上の各記事の信頼性を求めた．まず，編集履歴から，著者の編集履歴の維持割合を特定した．そして，その割合から著者の信頼度を算出した．

## 2.4 分析

### 2.4.1 Wikipedia における編集者の活動分析

山崎ら [6] は，編集者の編集プロセスをクラスタリングによって可視化を行うことにより，任意の 100 名の編者者の編集プロセスを分析した．この分析から，どのような編集者がどのように編集活動を行っているかを明らかにした．例えば広範囲・小編集編集タイプの編集者は，編集回数が少なく，同質でない記事に対して編集を行う編集者をさし，たまたま閲覧した記事の気になる箇所を修正するのみで，再編集は多くないという編集活動が推測される．

### 2.4.2　ウィキペディア記事閲覧回数の特徴分析

曽根ら [7] は，先行研究で余り扱われてこなかった，利用者の閲覧回数を表す，閲覧回数データの特徴分析を行った．その結果から，編集回数や編集ユニークユーザ数との相関がなく，よく編集される記事とよく閲覧される記事は異なることや，検索エンジンのヒット数との相関があることを確認した．

## 2.5　利用可能な Wikipedia の情報

これらの先行研究で利用されている，Wikipedia 固有の利用可能な情報について以下に整理する．本研究では特に，Wiki 構文の修飾記号を利用しセクションを抽出する．

**リンクネットワーク** Wikipedia のリンクによって作成されるネットワーク．

**編集履歴** 編集者の Wikipedia 編集履歴．

**閲覧回数** 利用者のページごとの閲覧数．

**Wiki 構文の修飾記号** リンク情報，セクション名などを表す修飾記号．

# 第3章　Wikipediaへの情報付与

本章では，教師あり機械学習を利用したWikipediaへの情報付与の手法，また情報付与に利用する技術について説明を行う．

## 3.1　問題設定

本研究では，Wikipediaから集めた「法則に関連するページ」に特化して研究を行う．これは，Wikipedia全体では範囲が大きすぎることと,「法則に関するページ」などに範囲を狭めることによって，効果の高い情報タグが想定しやすいためである．

抽出した法則に関連するページをセクションごとに分割し，セクションごとに特定の内容が書かれているか否かを，教師あり機械学習，または推定教師データに基づく機械学習によって判別し，情報タグの付与を行う．例えば，入力のセクションに歴史の内容が書かれているか否かを，教師あり機械学習によって判別し情報タグの付与を行う．入力セクションに歴史の情報が存在する場合は「:歴史」を，存在しない場合は「:無」を情報タグとしてセクションに付与する．

## 3.2　実験環境

実験には，2010年5月26日時点のWikipedia[8]を用い，認識性能が優れているSupport Vector Machine（以下SVM）[9]を実装しているTinySVM[10]，形態素解析を行うChaSen[11]を使用した．

## 3.3　Support Vector Machineへの適用

### 3.3.1　Support Vector Machine

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例のマージン（間隔）を大きくとるほど分類器の誤りが減少するという考えから，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う．一般的に上記の方法の他に,「ソフトマージン」と呼ばれる学習データにおいてマージンの内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，線形分離が不可能な問題に対応するために，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可能である．

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

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，

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

であり，また，各 $\alpha_i$ 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稿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．本稿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．

### 3.3.2 素性

本節では，素性 (解析に用いる情報) について説明する．本研究では素性に，文章中の名詞情報，年代情報の有無が利用可能である．名詞情報は，入力文章に対し ChaSen による形態素解析を行い，出力された名詞を抽出し素性として付与する．年代情報は，入力文章中に 4 桁あるいは 3 桁の数字，2 桁の数字と世紀，3 桁の数字と世紀が存在している場合，年代情報ありの素性を，それ以外の場合年代情報なしの素性を付与した．表 3.1 に素性を付与した例を示す．表において「名詞:X」は X という名詞が文章中に出現したことを意味する素性である．また,「年代 : あり」は，文章中に年代情報があることを意味する素性であり,「年代:なし」は，文章中に年代情報がないことを意味する素性である．

表 3.1: 素性例

| 本文 | 素性 |
|---|---|
| ==概論==ピーターの法則は、<br>「全ての有効な手段は、順次<br>さらに困難な応用に適用され、<br>やがては失敗する。」 | 年代:なし 名詞:概論 名詞:ピーター 名詞:法則<br>名詞:全て 名詞:有効 名詞:手段<br>名詞:困難 名詞:応用 名詞:適用<br>名詞:失敗 |
| <title>心理効果</title><br>”’心理効果”’（しんりこうか）とは、<br>心理がもたらす、様々な実際的・<br>具体的な効果、影響のこと。 | 年代:なし<br>名詞:心理 名詞:効果<br>名詞:心理 名詞:様々 名詞:実際 名詞:的<br>名詞:具体 名詞:的 名詞:効果 名詞:影響 名詞:こと |
| ==構造論への発展==1852 年に<br>エドワード・フランクランドは<br>有機金属化合物を合成し、<br>その際に有機金属化合物<br>中に含まれるアルキル基の数が<br>対応する金属ハロゲン化物や<br>水素化物のハロゲンや水素の<br>数と同じになることに気づいた。 | 年代:あり 名詞:構造 名詞:論 名詞:発展 名詞:年<br>名詞:エドワード 名詞:フランク 名詞:ランド<br>名詞:有機 名詞:金属 名詞:化合 名詞:物 名詞:合成<br>名詞:際 名詞:有機 名詞:金属 名詞:化合 名詞:物<br>名詞:中 名詞:基 名詞:数<br>名詞:対応 名詞:金属 名詞:ハロゲン 名詞:化物<br>名詞:水素 名詞:化物 名詞:ハロゲン 名詞:水素<br>名詞:数 名詞:こと |

## 3.4 法則ページ抽出方法

本研究で使用するコーパスを作成するために，法則名, 法則内容の抽出を Wikipedia から行う方法を示す.

### 3.4.1 Wikipedia の構成

Wikipedia は，Wiki と呼ばれる Web ブラウザを利用することによって，Web ページの編集等が行えるシステムを利用して作成されている．Wiki は Wiki 文法で記述され，Wikipedia で使用されている Wiki 文法には様々なものがあるが，その中で本研究で利用した文法について次に示す．リンク情報を表すタグは法則名抽出に利用され，タイトル，ページを表すタグは，法則内容抽出に利用し，セクション名を表すタグは，抽出したページの分割及びセクション名の抽出に利用した.

[BBB]……BBB というページへのリンク情報を表す.

<title>……ページのタイトル名の始まりを表す.

</title>……ページのタイトル名の終わりを表す.

<page>……ページの始まりを表す.

</page>……ページの終わりを表す.

==……セクション名の始まり，または終わりを表す.

### 3.4.2 法則名抽出

Wikipedia ではリンクが存在する単語は括弧で挟まれている．リンクが存在する単語は，その単語のページが存在している可能性が高いため，単純に単語を抽出するよりも，効率の良いページの抽出が可能になる．括弧で囲まれた単語の中から，正規表現にマッチする文字列を得ることによって，法則名を抽出する. 括弧で囲まれた単語のうち，[分布][分類][泳動][原理][現象][数][効果][公理] [理論][収差][定理][転位][予想][法][価][律][線][説][式][則] と末尾が一致したものを抽出し，その中から法則名でないものを人手で削除する.

### 3.4.3 法則内容抽出

抽出した法則名がタイトルとなっているページから，Wikipediaのタグを利用して法則内容を抽出する．<title>のタグから</page>のタグまでを1つのページとして取りだす．

### 3.4.4 法則内容の分割

Wikipediaから抽出した法則内容をセクション名の個所で文章を分割し、分割された文章の各部分を、各セクションの文章として抽出する．冒頭のセクションにはセクション名が存在しないため，タイトルで分割を行う．分割前のWikipediaの例を図3.1，分割後の例を図3.2に示す．

<title>グラフ理論</title>
<id>991</id>……
”’グラフ理論”’（グラフりろん、Graph theory）は……
== グラフの例 ==
乗り換え案内図: 前節の通り……
== グラフ理論の起源 ==
グラフ理論は、[[1736年]]……
=== 有向グラフ ===
集合V,Eと、Eの元に……
=== 無向グラフ ===
P(”V”) を ”V” の [[冪集合]] とする……
=== 頂点と辺 ===
グラフの頂点 [英 vertex; pl. vertices] 集合は……
=== 重みつきグラフ ===
グラフの辺に”’重み”’（”’コスト”’）が付いているグラフを、”’重み付きグラフ”’ ……
=== 接合と隣接 ===
辺の両端の点を”’端点”’ [英 endpoint] といい……
=== 距離と直径 ===
2頂点間の最短経路における辺数を距離 [英 distance ] と呼ぶ……
<\page>

図 3.1: 分割前データ

<title>グラフ理論</title>……グラフ理論……
== グラフの例 ==乗り換え案内図前節の通り……
== グラフ理論の起源 ==グラフ理論は、1736 年、……
=== 有向グラフ ===集合 V,E と、……
=== 無向グラフ ===P(V) を V の [[冪集合]] とする……
=== 頂点と辺 ===グラフの頂点 [英 vertex; pl. vertices] 集合は……
=== 重みつきグラフ ===グラフの辺に重み（コスト）が付いているグラフを……
=== 接合と隣接 ===辺の両端の点を端点 [英 endpoint] といい……
=== 距離と直径 ===2 頂点間の最短経路における辺数を距離 ……

図 3.2: 分割後データ

### 3.4.5 情報タグ候補の作成方法

利用者は，システムが出力した情報タグ候補の中から，付与したいタグの種類を選ぶ．情報タグの候補は，学習データの量が多いほど評価値の高い分類器が生成できると予測されるため，Wikipedia のセクション名の頻度の高い順に利用者に推薦される．まず，Wikipedia の「＝＝」のタグ情報を利用し，このタグで囲まれているセクション名を．そして抜き出したセクション名の頻度を調査し，頻度順に並べ替えを行うことによって情報タグの候補を作成する．

## 3.5 情報タグ付与方法

セクションに情報タグを付与手法を示す．

### 3.5.1 手作業で教師データを利用した SVM

手作業で教師データを作成し，SVM によって情報の有無を判断することによって，情報タグを付与する方法である．教師データの作成方法は，第 3.6.1 節に示す．

### 3.5.2 推定教師データを利用した SVM

手作業でデータに分類先を付与する作業を省略して作成する教師データ [1] を本研究では推定教師データと呼ぶ．Wikipedia から自動的に教師データを作成し，SVM によって

歴史情報の有無を判断することによって，情報タグを付与する方法である.「：歴史」の情報タグを付与する場合の推定教師データの作成方法を，第 3.6.2 節に示す.

### 3.5.3　パターンマッチング

年代情報を含むセクションに対して情報タグ「:歴史」を付与する方法である．年代情報を含むセクションとは，正規表現 [0-9][0-9][0-9][0-9]，[0-9][0-9][0-9]，[0-9][0-9] 世紀，[0-9] 世紀とマッチする表現を含むセクションのことである.

### 3.5.4　セクション名利用

Wikipedia につけられているセクション名を利用する方法である．例えば,「：歴史」の情報タグを付与する場合，歴史とセクション名がすでについているセクションに情報タグ「:歴史」を付与する.

### 3.5.5　全てに情報タグを付与

全てのセクションに情報タグを付与する方法である.「：歴史」の情報タグを付与する場合，すべてのセクションに対し情報タグを付与する.

## 3.6　教師データ作成手法

### 3.6.1　手作業による教師データの作成

抽出した 20,001 個のセクションからランダムで 1,000 個取りだし，その取り出したセクションのそれぞれに対して，歴史の情報を含んでいるか否か示すタグを手作業で付与することにより，教師データを作成する.

### 3.6.2　推定教師データの作成

手作業でデータに分類先を付与する作業を省略し，推定教師データを作成する．例えば，情報タグ「：歴史」を付与するための教師データを作成する場合，Wikipedia には,「歴史」というセクション名が存在しているページがある．そういったページでは，セクション名が「歴史」のセクションは歴史の情報を含み, 他のセクションには，歴史の情

報が含まれていない可能性が高い．推定教師データは，その性質を利用するものであり，推定教師データは以下の手順で構築される．

**手順１** 法則関係のページからセクション名に「歴史」があるページを抽出する．

**手順２** 抽出したページについて，ページごとにセクション名「歴史」のセクションをクラス「歴史」に，それ以外のセクションをクラス「無し」として教師データを作成する．

**手順３** 書きかけのセクション等を削除する．

表 3.2 に熱力学のページの各セクションの分類例を載せる．

表 3.2: 分類例

| セクション名 | 分類先 |
|---|---|
| 熱力学 | 無し |
| 目次 | 無し |
| 歴史 | 歴史 |
| 熱力学の法則 | 無し |
| より百科事典的な説明 | 無し |
| 熱力学的系 | 無し |
| 基本法則からの発展と応用 | 無し |
| 非平衡熱力学 | 無し |
| 参考文献 | 無し |
| 関連書籍 | 無し |

表のように，セクション名が「歴史」であったセクションは分類先が「歴史」となり，それ以外は「無し」となる．

# 第4章　実験

本章では，前章で述べた方法での情報タグ付与を行う．

## 4.1　事前準備

本実験で使用するコーパスを作成するため，Wikipedia からの法則のページの抽出を行った．

### 4.1.1　法則名抽出

第 3.4.2 節に示した方法によって，法則名（12,924 個）を獲得した．表 4.1 に獲得した法則名例を，表 4.1.1 に人手により排除した法則名例を載せる．[線][説][式][則] 等一文字のパターンでは，他の言葉とマッチしてしまう可能性が高いことがわかった．

表 4.1: 獲得法則名例

| 法則の名前 |
|---|
| アムダールの法則 |
| ジムロート転位 |
| ヒルベルト曲線 |
| ジップの法則 |
| ホイップル効果 |

表 4.2: 排除した法則名例

| 誤った法則の名前 |
|---|
| NTT 出版・公式 |
| 和歌山県道 24 号御坊由良線 |
| フィラデルフィアでの演説 |
| 恋のミクル伝説 |
| 伊藤政則 |

### 4.1.2 法則内容抽出

第 3.4.3 節に示した方法によって，法則に関するページ (5,061 ページ，20,001 項目) を抽出した.

### 4.1.3 法則内容分割

第 3.4.4 節に示した方法によって，抽出した法則に関するページを 1 行 1 セクションの形式に分割した．分割結果のうち,「行列の階数」のページを表 4.3 に「モンテカルロ法」のページを表 4.4 に示す.

表 4.3: 抽出例

| |
|---|
| <title>行列の階数</title>…… 線型代数学において、行列 …… |
| ==定義==行列の階数について、文献によっては列ベクトルの…… |
| ==性質==”A” を”m” ×”n” 行列とする。また、”f” を…… |
| ==階数の計算==例えば、行列は、[[行列の基本変形—…… |
| ==線型写像の階数==”V, W” をそれぞれ…… |
| ===次元定理===”V, W” を有限次元ベクトル空間とし…… |

表 4.4: 抽出例 2

| |
|---|
| <title>モンテカルロ法</title>……モンテカルロ法…… |
| ==計算理論==[[計算理論]] の分野において、モンテカルロ法と…… |
| ==準モンテカルロ法==乱数ではなく、[[一様分布列]] …… |
| ==[[数値積分]]==[[数値解析]] の分野に於いてはモ…… |
| ==機械学習==[[機械学習]] の分野におけるモンテカルロ法とは…… |
| ==統計学==統計学におけるモンテカルロ法の 1 つとして、…… |
| ==乱数の選択==モンテカルロ法では状況に応じた乱数…… |
| ==精度==また、精度の良い結果を得るためには多く…… |
| ==関連項目==* [[ブートストラップ法]]…… |
| ==参考文献==* Jan van Leeuwen 編…… |

## 4.2 情報タグ候補の決定

第 3.4.5 節に示した方法によって，抽出した法則のページから情報タグの候補を獲得した．本研究では，十分な推定教師データが作成できるように，頻度の高いセクション名を情報タグの候補とする．表 4.5 に頻度順のセクション名とその頻度を示す．この表の上位の頻度を持つセクション名が，情報タグの候補となる．

本研究では，この表の上位頻度のセクション名の中でも，法則のページの主要な分類であろう [歴史][証明][例][定義] に関して情報タグの付与を行い，評価を行うこととした．

表 4.5: 情報タグ候補

| セクション名 | 頻度 |
|---|---|
| 関連項目 | 1755 |
| 外部リンク | 758 |
| 参考文献 | 716 |
| 概要 | 534 |
| 脚注 | 407 |
| 歴史 | 159 |
| 定義 | 121 |
| 例 | 119 |
| 構成 | 76 |
| 概説 | 64 |
| 証明 | 57 |
| 出典 | 55 |
| … | … |

## 4.3 教師データ作成

第 3.6.2 節の方法を利用し，[歴史][証明][例][定義] の情報タグ付与に使用する推定教師データの作成を行った.「歴史」の情報タグについては，手作業で作成された教師データとの比較を行うため教師データを作成した．表 4.6 にそれぞれの推定教師データ数とその内訳を，表 4.7 に教師データ数とその内訳を示す.

表 4.6: 推定教師データ数

| セクション名 | 推定教師総数 | 情報あり |
|---|---|---|
| 歴 | 1,477 | 306 |
| 例 | 868 | 117 |
| 定義 | 991 | 124 |
| 証明 | 305 | 69 |

表 4.7: 教師データ数

| セクション名 | 教師総数 | 情報あり |
|---|---|---|
| 歴 | 1,000 | 129 |

## 4.4　セクション名の付与結果

第 4.3 節で作成された推定教師データで学習した SVM，教師データで学習した SVM，パターンマッチング，セクション名利用，全て歴史に分類の 5 手法によって，法則のページの 20,001 セクションに [歴史][証明][例][定義] のセクション名を付与した．教師データで学習した SVM によって，歴史の情報タグを付与したグリセミック指数のページを図 4.1 に示す．歴史の内容が書かれているセクションのセクション名「グリセミック指数」「低インシュリンダイエット」に「：歴史」が付与されていることがわかる．

**グリセミック指数:歴史**
グリセミック指数 (glycemic index) とは、……1981 年に……
**測定基準**:**無**[編集]
食品の炭水化物 50 グラムを摂取した際の血糖値上昇の度合いを……
**健康**:**無**[編集]
いくつかの研究結果で GI 値の低い食べもので食生活を組み立てた……
**低インシュリンダイエット**:**歴史**[編集]
米国発祥のダイエット方法（シュガーバスター）に……
**脚注**:**無**[編集]
^ Jenkins DJ et al.……
**関連項目**:**無**[編集]
血糖値

図 4.1: 付与結果例

# 第5章　評価

本章では，第 4 章で説明した各手法の比較評価を行う．

## 5.1　評価値の計算

### 5.1.1　F 値の計算

精度は F 値で比較を行う．F 値とは適合率と再現率の調和平均であり，(5.1) の式を用いて算出する．適合率はシステム出力の正解率，再現率は問題に対する取りこぼしの指標であり，適合率は (5.2)，再現率は (5.3) の式で表される．

$$F\text{値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \tag{5.1}$$

$$\text{適合率} = \frac{\text{システムの正解数}}{\text{システムの出力数}} \tag{5.2}$$

$$\text{再現率} = \frac{\text{システムの正解数}}{\text{テストデータ中の正解数}} \tag{5.3}$$

### 5.1.2　交差検定

交差検定（cross-validation）は，教師データを n 分割し 1 個をテストデータ，残りの n-1 個を教師データとして学習を行い，分割された教師データが全て 1 回テストデータとして用いられるのように繰り返して精度を求める方法である．本実験では，手作業で作成された教師データの SVM に対して，10 分割の交差検定によって精度を求める．

### 5.1.3 ブートストラップ法

ブートストラップ法は，リサンプリング法の 1 つであり，標本からのリサンプリングを繰り返すことにより母集団の性質を推定する方法である．リサンプリング数は 10,000 回程度が一般的である．また，リサンプリングを B 回行い，仮説が支持された数が C 回とするとブートストラップ確率は (5.4) 式で表される．

$$\text{ブートストラップ確率} = \frac{C}{B} \tag{5.4}$$

下平は [12] これが 1 に近いほど仮説はもっともらしく，0 に近いほど仮説は疑わしいと述べている．

## 5.2 情報タグ「歴史」の評価

推定教師データを利用した SVM で生成された分類器の歴史に関する F 値を，他の手法と比較する．

### 5.2.1 比較手法

各手法の説明は以下の通りである．

**手法１** 推定教師データ（教師データ数 1,477 件うち歴史 306 件) を利用した SVM.

**手法２** 手作業教師データ（教師データ数 1,000 件うち歴史 129 件) を利用した SVM.

**手法３** 歴史とセクション名がすでについているものをセクション名「歴史」に分類する，セクション名利用による方法.

**手法４** 「年代」を含むセクションを「歴史」と分類するパターンマッチングによる方法.

**手法５** 全て「歴史」に分類する方法.

すべての手法でテストデータは 1,000 件の教師データを用いる．ただし，手法 2 ではその 1,000 件の教師データで 10 分割交差検定で F 値を求める．

### 5.2.2 比較結果

表 5.1 に各手法の F 値を示す．括弧内は教師データの種類であり，教師データを使用しない場合は無しになっている．

表 5.1: 手法の比較

| 手法 (教師) | F 値 | 再現率 | 適合率 |
|---|---|---|---|
| SVM(手作業) | 0.612 | 0.806 | 0.493 |
| SVM(推定) | 0.524 | 0.512 | 0.537 |
| セクション名利用 (無し) | 0.129 | 0.069 | 0.818 |
| パターン（無し） | 0.554 | 0.937 | 0.394 |
| 全て歴史（無し） | 0.220 | 1.000 | 0.129 |

以上から，手作業の教師データを利用した SVM が最も精度が高く，推定教師データを利用した SVM もパターンによる手法に近い精度が得られていることがわかる．推定教師を利用した SVM と，パターンによる手法に関しては値が近いため有意差検定を行う．

### 5.2.3 有意差検定

推定教師データを利用した SVM と，F 値が近いパターンによる手法をブートストラップ法によって検定した．重複を許して手作業で作られた教師データから 1,000 件を抜き出し，それをテストデータとして利用する．テストデータを各手法でセクション名を付与し，F 値を比較する．これを 10,000 回繰り返し，ブートストラップ確率が 0.95 以上の F 値が存在する場合に有意差があるとする．検定結果を表 5.2 に示す．推定教師データを利用した SVM と，パターンによる手法とでは，有意差が無いため，ほぼ同等な精度が得られていることがわかる．

表 5.2: ブートストラップ法

| | 推定教師 | パターン |
|---|---|---|
| ブートストラップ確率 | 0.2068 | 0.7932 |

## 5.3　証明，例，定義の評価

### 5.3.1　評価値算出方法

[証明][例][定義] のセクションに対して，手作業で作成する教師データを用意していないため，以下の方法で F 値を求める．まず，法則のセクション 20,001 件に推定教師データを利用した SVM を利用して分類を行う．そして，情報ありと分類されたデータ 30 件，情報なしと分類されたデータ 30 件をランダムで抜きだす．その正解数を手作業で確認し適合率を求める．出力データの情報あり，情報なしに分類された各総数を調べ適合率から，入力データ中の実際の情報あり，情報なしの総数を予測し再現率を求める．

### 5.3.2　評価結果

[証明][定義][例] 各分類ごとの情報ありと分類されたデータ 30 件，情報なしと分類されたデータ 30 件中の正解数を調査した．表 5.3 に証明の，表 5.4 に定義の，表 5.5 に例の結果を示す．

表 5.3: 評価：証明

| | 評価 |
|---|---|
| 証明 | 3 |
| 無 | 30 |

表 5.4: 評価：定義

| | 評価 |
|---|---|
| 定義 | 14 |
| 無 | 24 |

表 5.5: 評価：例

| | 評価 |
|---|---|
| 例 | 10 |
| 無 | 25 |

### 5.3.3 F 値の算出

第 5.3.1 節の方法で，[証明][例][定義] の推定教師データによって生成される分類器の，[証明][例][定義] の F 値を全てそのセクションに分類する方法と比較した．表 5.7 に結果を示す．すべての分類に対して F 値が上昇しているが，各分類ごとの差は小さいため，各分類に対してブートストラップ法によって有意差検定を行う．

表 5.6: F 値の比較

| 分類 | 予測 F 値（SVM） | 予測 F 値（全て歴史） |
|---|---|---|
| 証明 | 0.182 | 0.012 |
| 例 | 0.366 | 0.346 |
| 定義 | 0.543 | 0.422 |

表 5.7: 再現率適合率の比較

| 分類 (手法) | 再現率 | 適合率 |
|---|---|---|
| 証明 (SVM) | 1.000 | 0.100 |
| 証明 (全て歴史) | 1.000 | 0.006 |
| 例 (SVM) | 0.406 | 0.333 |
| 例 (全て歴史) | 1.000 | 0.209 |
| 定義 (SVM) | 0.440 | 0.466 |
| 定義 (全て歴史) | 1.000 | 0.260 |

### 5.3.4 有意差検定

推定教師データを利用した SVM と，全て歴史に分類する手法をブートストラップ法によって検定した．情報ありと分類されたデータ 30 件から，重複を許して 30 件ランダムに抜きだして各手法と F 値を比較する．これを 10,000 回行い，ブートストラップ確率が 0.95 以上の手法に有意差があるとする．検定結果を表 5.8 に示す．[証明] は有意差あり，その他は有意差なしであることがわかる．

表 5.8: ブートストラップ法 2

| | 推定教師 | 全て歴史 |
|---|---|---|
| 証明 | 0.9585 | 0.0415 |
| 例 | 0.5941 | 0.4059 |
| 定義 | 0.6594 | 0.3406 |

## 5.4 教師データと推定教師データの組み合わせ

教師データと推定教師データの組み合わせによる F 値の変化を調査した．

### 5.4.1 Stacking

Stacking[13] は，ある教師データを利用し機械学習を行い，その推定結果を素性として他の教師データに追加し，学習を行う方法である．Stacking を行い，推定教師データを利用した SVM の推定結果を，手作業で作成した教師データに素性として追加し SVM による機械学習を行った．結果を表 5.9 に示す．手作業で作成した教師データを利用した SVM の F 値（0.612）と比べ値が低下しているため，効果的でない．

表 5.9: スタッキング

| | F 値 | 再現率 | 適合率 |
|---|---|---|---|
| スタッキング | 0.577 | 0.752 | 0.469 |

### 5.4.2 推定教師データと教師データの合成

教師データに推定教師データを加え学習を行った．結果を表 5.10 に示す．手作業で作成した教師データを利用した SVM の F 値（0.612）と比べ値が低下しているため，効果的でない．

表 5.10: 教師＋推定教師

| | F 値 | 再現率 | 適合率 |
|---|---|---|---|
| 教師＋推定教師 | 0.207 | 0.667 | 0.123 |

## 5.5 教師データ数ごとの評価値

教師データ，推定教師データ数の変化による F 値の変化を調べた．

### 5.5.1 教師データ

1,000 件の教師データをクロスバリデーションの分割数を変化させることによって，教師データ数ごとの歴史に関する F 値を調査した．結果を表 5.11 に示す．教師データ数の増加による F 値の上昇を確認できる．

表 5.11: 教師データ数ごとの評価

| 教師数 | F 値 |
|---|---|
| 1,000 | 0.612 |
| 500 | 0.586 |
| 250 | 0.572 |
| 125 | 0.547 |
| 100 | 0.526 |

### 5.5.2 推定教師データ

1,477 件の推定教師データの中から，ランダムで指定した数を重複を許してリサンプリングを行い，教師データを作成する．リサンプリングは 10 回行い平均 F 値を抽出する．教師データ数ごとの歴史に関する F 値を調査した．結果を表 5.12 に示す．推定教師データ 250 件が最も高い．リサンプリング回数が少なく数値が安定しないためと考えられる．

表 5.12: 推定教師データ数ごとの評価

| 推定教師数 | F 値 |
|---|---|
| 1,477 | 0.524 |
| 1,000 | 0.532 |
| 500 | 0.54 |
| 250 | 0.562 |
| 125 | 0.49 |
| 100 | 0.411 |

# 第6章　考察

本章では，第 4 章，第 5 章の結果から考察を行う．第 6.1 節では，教師データ，F 値，情報タグ付与結果から支援の必要性を示す．第 6.2 節では，推定教師データと他手法との比較，推定教師データを利用した SVM によって付与した情報タグごとの比較を行うことにより，推定教師データの特徴を明らかにする．第 6.3 節では，各手法の特徴について述べ，第 6.4 節で各手法の組み合わせによる F 値向上の可能性について考察する．

## 6.1　支援の必要性

### 6.1.1　教師データからの考察

表 6.1 に手作業で作成した教師データ中の，歴史の情報を含む 129 件のうち，セクション名が歴史の件数，セクション名に歴史を含む件数とセクション名に歴史を含まない件数を示す．

表 6.1: 歴史の情報を含むセクションの内訳

| | 個数 |
|---|---|
| セクション名が歴史 | 9 |
| セクション名に歴史を含む | 11 |
| セクション名に歴史を含まない | 118 |

教師データ中の情報タグ「:歴史」が付与されたセクションの中でセクション名に歴史が含まれていないものを調査した．結果の一部を表 6.2 に示す．最も頻度が高い「概要」からは歴史情報の有無が判断できない．他のセクション名にも同様なことが言える．歴史情報の有無が判断できるセクション名が存在していれば，情報タグを付与する効果は低いが，そのようなセクション名は少ないため，本研究の歴史の情報タグ付与の効果が大きいことがわかる．

表 6.2: 情報タグ「:歴史」の内訳

| セクション名 | 頻度 |
|---|---|
| 概要 | 8 |
| 略歴・概要 | 1 |
| 理論の拡張と応用 | 1 |
| 来歴 | 1 |
| 有力な出資スポンサー企業 | 1 |
| 報道規制 | 1 |
| 表記の一覧 | 1 |
| 反論 | 1 |
| 発端 | 1 |
| 八月革命説に対する批判 | 1 |
| 白鳥の首フラスコ実験 | 1 |
| 背景 | 1 |
| 年紀法の定着と消滅 | 1 |
| 日本国内の状況 | 1 |
| 内生的成長モデル | 1 |
| 特徴 | 1 |
| 導出 | 1 |
| 当時のイングランドの状況と「堕落した聖職者」問題 | 1 |
| 土地の剥奪 | 1 |
| 展開 | 1 |
| 鉄杭除去運動 | 1 |
| 哲学者の回答 | 1 |
| 定数の値 | 1 |
| 定義 | 1 |
| 前段階 | 1 |

セクション名に歴史を含まないが，セクション中に歴史の情報が存在しているものを次に示す．図 6.1 に冒頭のセクションに，歴史の情報が存在している場合の例を示す．一行目のビリアル定理がページ名であり，太字が歴史情報である．歴史の情報が冒頭のセクションに存在している場合，セクション名がつけられていないため，いずれかの手法で情報タグを付与する必要がある．

ビリアル定理
出典: フリー百科事典『ウィキペディア（Wikipedia）』
移動： 案内, 検索
ビリアル定理 (Virial theorem) とは……
**……によって定義される値で、1870 年クラウジウスが命名した。**

図 6.1: 冒頭のセクション

図 6.2 に歴史と近い意味を持つセクション名の例を示す．一行目の背景がセクション名であり，太字が歴史情報である．背景は歴史と近い意味をもつセクション名であり，セクション内に歴史の情報を含む可能性が高い．現状でも利用者の判断でセクション内の歴史情報の有無を判断できる可能性が高い．しかし，情報タグを付与することによってさらに明確に情報の有無を確認できることが予測される．

背景**核の説は最初に 1836 年にオーギュスト・ローランによってその一部が示され**
翌年博士論文の中で全体が発表された。 当時はイェンス・ベルセリウスの
電気化学的二元論とそれによって立つユストゥス・フォン・リービ……

図 6.2: 歴史と類義語のセクション名の例

図 6.3 に最も情報タグの効果高いと考えられる例を示す．一行目のコペルニクスの地動説がセクション名であり，太字が歴史情報である．セクション名からは，歴史情報の有無がまったく予測できず，支援の必要がある．

コペルニクスの地動説
カトリック教会の司祭であったコペルニクスは、この誤差に着目した。
彼は地動説を新プラトン主義の太陽信仰と……
コペルニクスは没年にあたる 1543 年、
思索をまとめた著書『天体の回転について』を刊行した。
……創始者とされる理由である。またこの業績について、ガリレオ
・ガリレイから「太陽中心説を復活させた」と評された。

図 6.3: 最も効果のある例

### 6.1.2 F 値からの考察

表 5.1 の結果から，セクション名利用の F 値が他の手法に比べて圧倒的に低いことがわかる．このことから既存のセクション名だけでは，効率のよい情報収集ができないため支援の必要性が示された．セクション名利用の F 値が低い理由については，1 つのセクションに 1 つのセクション名が付けられているため，セクション中に複数の情報が入っている場合，適切なセクション名をつけることが難しいということがあげられる．例えば，図 6.4 の「全か無かの法則」のページ内の概要のセクションでは，最後の行で発見年つまり歴史の事柄が補足的に書かれているため，概要と歴史の情報が混在してしまっている．

概要 [編集]
全か無かの法則は、筋線維や神経線維に加えた刺激が弱いと反応しないが、
限界値（閾値）に達すると最大限度に反応するといったことを示した法則で
ある。閾値を越えた刺激を与えたとしても、線維の反応状態は変わらない。
但し筋肉などを刺激する場合は、刺激に反応する線維の本数が多く、線維一
本一本の閾値が異なるため、筋肉全体で見れば刺激への反応はこの法則に従
わない。
この法則は、1871 年に H・P・ボウディッチ（Henry Pickering Bowditch）が行った、
カエルの心臓を用いた実験により提唱された。

図 6.4: 情報混在例

### 6.1.3 情報タグ付与結果からの考察

Wikipedia の法則に関連するページ 20,001 セクション対して，手作業で教師データを用いた SVM によって「：歴史」の情報タグを付与した．結果を表 6.3 に示す.『「：歴史」タグ数』は，法則に関連するページに付与された情報タグ「：歴史」の数,『セクション名「歴史」総数』はセクション名「歴史」の数,『「：歴史」＋「歴史」数』は情報タグ「：歴史」が付与されたセクション名「歴史」の数である．もとからセクション名「歴史」と記載されているセクションには，約半数，情報タグ「：歴史」をふることができていることがわかる．表 6.4 に再現率，適合率から，情報タグの正解数，法則に関連するセクション 20,001 個中の情報タグの正解数と歴史の情報を含むセクション数を予測した値を示す．これらの結果から，既存のセクション名を手掛かりに歴史情報を探した場合，全体の 14 %程度しか得ることができないため，効率の良い情報収集が困難であることが明らかになった．

表 6.3: 情報タグ付与結果

| | 個数 |
|---|---|
| セクション総数 | 20,001 |
| 「：歴史」タグ数 | 4,385 |
| セクション名「歴史」総数 | 306 |
| 「：歴史」＋「歴史」数 | 153 |

表 6.4: 法則関連ページ予測数

| | 予測数 |
|---|---|
| 正解「：歴史」タグ数 | 2,161 |
| 歴史の情報を含むセクション | 2,680 |

## 6.2 推定教師データ

### 6.2.1 教師データとの比較

前章で，推定教師データは，教師データよりも F 値が低いことがわかった．これは，第 6.1 節に説明した理由によるもので，自動的に作成された推定教師データは，教師データ

よりも誤りの情報が多いからではないかと考えられる．例えば，フラッシュ法では，初めのセクションに歴史情報が存在しているが，歴史というセクション名ではないため推定教師データでは，歴史のセクションとは扱われていない．図 6.5 にフラッシュ法のページを示す．

フラッシュ法
出典: フリー百科事典『ウィキペディア（Wikipedia）』
移動： 案内, 検索
フラッシュ法（フラッシュほう、英: Frasch process）は……2002 年時点では……
概要 [編集]
硫黄鉱床を覆う岩にドリルにより穴をあけ、二重にした同心円状のパイプを通……
歴史 [編集]
1867 年に、アメリカ合衆国ルイジアナ州カルカシュー教区の岩塩ドームで……
脚注 [編集]
^ D'Arcy Shock (1992) "Frasch sulfur mining,……

図 6.5: フラッシュ法

### 6.2.2 パターンマッチング手法との比較

前章で推定教師データを利用した SVM は，パターンマッチング手法とほぼ同等の F 値であることを述べた．しかし，これらの手法は適合率，再現率の値に大きな違いがある．表 5.1 から，次のことがわかる．パターンマッチング手法は，再現率が高く，適合率が低いことから，テストデータ中の歴史のセクションを取りこぼしを少なく分類できるが，その分誤りも多く出力してしまうことがわかる．一方推定教師データを利用した SVM は，パターンマッチング手法と比べて適合率が高く誤りの出力は少ないが，再現率が低いため取りこぼしの数が多くなっている．再現率重視ならパターンマッチングの手法，適合率重視ならば推定教師データを利用した SVM を利用すると効果的である．

### 6.2.3 情報タグごとの比較

表 4.6 と表 5.7 から推定教師数の増加に比例して F 値が上昇していることがわかる．[証明] の F 値が他と比べて極端に低いのは，推定教師データ数が少ないことと，今回の素性が，[証明] の分類を行うのに適切で無かったためだとが考えられる．[証明] のセクション

は，名詞情報が少なく現状の素性だけでは特定が難しい．[証明] のセクションの素性には,「なぜなら」等の接続語,「証明された。」等の文章の語尾情報などの素性が適していると予測される．図 6.6 に証明のセクション例を示す．

証明 [編集]
二項定理から、数学的帰納法を用いて、……という形の命題として証明する。
1. (m + 1)p という式を展開することを考える。これは公式により、
mp + pC1mp-1 + pC2mp-2 + … + pCp-1m + 1
となる。
2. ここで両端の項以外はすべての項に……れないからである。
3. すると、両端の項以外は p で割り切れる。……と等しいことになる。
4. ここで、m = 1 とする。2p = (1 + 1)p を……正しいことが証明された。
5. a に関する帰納法で示すために……すべての a について命題は成立することになる。
6. a = 0, 1 の場合は命題の成立することが自明。
（証明終わり）

図 6.6: 証明のセクション例

### 6.2.4 利点と欠点

以上の考察から推定教師データの利点として，次のことがわかった．

**利点１** 教師データの作成等人手作業を大幅に短縮することができる

**利点２** 教師データを利用したSVMを除く他の手法と比べて，高い精度が出る場合が多い

また，欠点としては次のことがわかった．

**欠点１** 手作業で作られた教師データよりも誤りを含みやすい

**欠点２** 推定教師データ数が少ない場合，F 値が極端に落ちる

## 6.3 各手法の特徴

表 6.5 に各手法の特徴を示す．作成コストは，パターンや教師データの作成コストである．手作業で作成した教師データを利用した SVM は，教師データ作成コストは高いが F 値が高い．推定教師データは，F 値は手作業教師データよりも低いが，作成コスト

がほとんど無く，F 値もそれなりの値を出す．作成コストを払えない場合に役立つ手法である．パターンマッチングは，再現率が高いため，取りこぼしをなるべく支度ない場合に効果的な手法である．

表 6.5: 手法の特徴

| 手法 (教師) | F 値 | 再現率 | 適合率 | 作成コスト |
|---|---|---|---|---|
| SVM(手作業) | 高 | 高 | 中 | 大 |
| SVM(推定) | 中 | 中 | 中 | 無 |
| パターン（無し） | 中 | 高 | 低 | 中 |
| セクション名利用 (無し) | 低 | 低 | 高 | 無 |
| 全て歴史 | 低 | 高 | 低 | 無 |

## 6.4 手法の組み合わせ

### 6.4.1 他手法と教師あり機械学習の組み合わせ

各手法の特性を生かし組み合わせることによって，F 値の向上の可能性がある．例えば，適合率の高い方法でまず判定し，次に再現率の高い方法を利用することで，適合率を高めつつ，再現率の高い判定ができると考えられる．本研究での手法の組み合わせ手法としては，以下の方法が考えられる．まず，適合率の高いセクション名利用の手法によって正例と判断されたデータを正例としてテストデータから抜きだす．つぎに，再現率の高いパターンマッチングによる手法で，テストデータから負例と分類されたものを排除する．最後に残ったテストデータを入力とし，手作業で教師データを作成した SVM で分類を行う．

### 6.4.2 教師データと推定教師データ組み合わせ

第 5.4 節から，単純な教師データの組み合わせでは F 値の向上が困難であることがわかった．しかし，第 5.5 節からわかるように，手作業で作成された教師データが 100 件の F 値が，推定教師データ 1,477 件の F 値とほぼ同程度である．この場合であれば，F 値が同程度であるので，この二つの教師データを組み合わせることによって，F 値向上の可能性があると考えられる．

# 第7章　おわりに

本研究では，Wikipedia 利用者の支援のために，Wikipedia のセクションごとに歴史が存在しているか否かを表す情報タグの付与を行った．また，実際に Wikipedia のセクションに情報タグを様々な手法で付与し F 値を調査した．歴史の情報タグ付与実験でわかったことを以下に整理する．

1. 手作業で作成した教師データを利用した SVM は，F 値 0.612（再現率 0.511，適合率 0.536）で歴史の情報タグを付与することができた．

2. 推定教師データを利用した SVM は，F 値が 0.524（再現率 0.806，適合率 0.492）であり，手作業で作成した教師データを利用した SVM の F 値よりも低いが，パターンマッチングの F 値 0.554（再現率 0.937，適合率 0.394）とほぼ同程度の F 値であることがわかった．

3. セクション名利用による方法の F 値が極端に低い 0.129（再現率 0.069，適合率 0.818）であることから，既存のセクション名では情報が不足しており，セクション名に何らかの支援が必要であることが明らかになった．これは，本研究で行なった，手作業で作成した教師データを利用した SVM や推定教師データを利用した SVM やパターンマッチングの手法による，セクションへの情報付与が役立つことを意味する．

4. 全てを歴史と判定する方法は F 値が低く，手作業で作成した教師データを利用した SVM や推定教師データを利用した SVM や推定教師データを利用した SVM やパターンマッチングの手法などの情報タグ付与手法の利用が必要なことがわかった．

5. F 値のみならず，再現率，適合率に基く分析を行い，各手法の特徴を明らかにした．例えば，情報の取りこぼしなどを防ぎたいなど，適合率は低くても再現率を重視する場合は，パターンマッチングの手法が役立つこと，中程度の再現率，適合率でよいがコストの低い手法が利用したい場合は，推定教師データに基づく機械学習による手法が役立つことを明らかにした．

さらに，証明，例，定義の情報タグ付与を行った．この実験より以下のことがわかった．

1. 推定教師データを利用したSVMは，[証明]のF値0.182（再現率1，適合率0.1），[例]のF値0.366（再現率0.406，適合率0.333），[定義]のF値0.543（再現率0.44，適合率0.446）で情報タグを付与することができた．

2. 証明の推定教師データ数（305個），例の推定教師データ（868個），定義の推定教師データ（991個）から，F値は，推定教師データ数に比例する可能性があることがわかった．

さらに、Stacking，教師データと推定教師データの組み合わせ，教師データ数によるF値の変化，推定教師データ数によるF値の変化の追加実験を行うことで、以下のことがわかった。

1. 教師データ数が十分用意できる場合，Stackingは有効でないが，教師データ数が少ない場合，F値向上の可能性がある．

2. 教師データ数が十分用意できる場合，推定教師データとの併用は有効でないが，教師データ数が少ない場合，F値向上の可能性がある．

3. SVMのF値は教師データ数に比例しており，教師数が100件の場合のF値0.526が推定教師1,477件のF値とほぼ同値であることがわかった．

4. SVMのF値は推定教師データ数にほぼ比例していることがわかった．

今後の課題としては，現段階では実用に耐えられるF値ではないため，素性や手法の改良を行うことによって，F値を向上させる必要がある．特に素性に関しては，年代情報と名詞情報しか使用していないため，教師データ総数を増加させた場合，推定が困難になることが予測されるため，文末情報，文長等を増やす必要がある．今回，評価に関してはひとりで行ったため信頼性の低い評価となっている．今後は複数人で行い信頼性の高い評価を行いたい．また，本研究で付与した情報タグの有用性の評価を行いたい．

# 謝辞

本研究を進めるにあたり，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，村上仁一准教授には，統計的観点からの様々な御指導を頂きました．ここに深く感謝いたします．本研究を進めるにあたり，御指導を頂きました徳久雅人講師に心から御礼申し上げます．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] 藤井 敦, 三條場 旭彦: “Wikipedia を用いた用語説明のモデル化と事典的検索への応用”, 人工知能学会 第 20 回セマンティックウェブとオントロジー研究会資料, SIG-SWO-A803-01, pp.1-8, 2009.

[2] 隅田 飛鳥, 吉永直樹, 鳥澤 健太郎, 萬成賢太郎: “Wikipedia からの大規模な上位下位関係の獲得”, 言語処理学会　第 14 回年次大会　発表論文集, pp.796-772, 2008.

[3] 野田陽平, 清田陽司, 中川裕志: “Wikipedia カテゴリネットワークからの意外性のある関係性の抽出”, 第 21 回セマンティックウェブとオントロジー研究会, SIG-SWO-A901-04, pp.1-4, 2009.

[4] 中川 隆人, 古瀬 一隆, 陳 漢雄: “Wikipedia におけるミッシングリンクの自動発見手法”, 情報処理学会創立 50 周年記念（第 72 回）全国大会, pp.749-750, 2010.

[5] 鈴木 優, 金本 径卓, 川越 恭二: “Wikipedia の編集履歴に基く記事の信頼性導出”, 人工知能学会 第 20 回セマンティックウェブとオントロジー研究会資料, SIG-SWO-A803-09, pp.1-8, 2009.

[6] 山崎 由佳, 井庭 隆, 熊坂 賢次: “Wikipedia における編集者の活動分析”, 人工知能学会 第 21 回セマンティックウェブとオントロジー研究会資料, SIG-SWO-A901-02, pp.1-7, 2009.

[7] 曽根 広哲, 山名 早人: “ウィキペディア記事閲覧回数の特徴分析”, 人工知能学会 第 21 回セマンティックウェブとオントロジー研究会資料, SIG-SWO-A901-03, pp.1-5, 2009.

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

[9] 奥村 学, 高村大地: “言語処理のための機械学習入門”, コロナ社, 2010.

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

[11] ChaSen: http://chasen-legacy.sourceforge.jp/

[12] 下平 英寿: “ブートストラップ法によるクラスタ分析のバラツキ評価”, 統計数理, 50, pp.33-44, 2002.

[13] 村田 真樹，井佐原 均: “受け身／使役文の能動文への変換における機械学習を用いた格助詞の変換”, 情報処理学会 自然言語処理研究会, 2002-NL-149, pp.39-44, 2002.