概要

物事の変遷を知ることはその物事の知識を会得する時に重要であるが，人手では網羅的に収集するのが困難であり，かつ多大な労力を要する．そこで，変遷を知ることを自動で簡単に行いたい．

堀らの先行研究 [1] ではその手始めとして，論文のタイトル，著者名のデータを使用し，どの分野がどの分野から発生したか，どの研究者がどの研究者を教示していたかの関係 (いわゆる先生と弟子のような関係) を自動で推定した．

この研究は，問題点として学術分野間，師弟間という限定された変遷情報の種類についての抽出であったことが挙げられる．

本研究では，この問題を解決するために，まず大量の文からパターンに合致するものを取得することで，幅広い分野の変遷情報を取得した．この方法で 0.86 という高い F 値で変遷情報を抽出できた．更に，より性能高く変遷情報を抽出するために，パターンベース法に機械学習 (SVM,ME) を追加した．実験の結果，SVM では F 値 0.91，ME では F 値 0.89 で変遷情報を抽出できた．ただし，ここでの F 値はパターンに基づく方法で抽出できた正しい変遷情報の個数に相当するものを再現率の分母にして算出したものである．詳しくは 4.2.2 節において説明する．本研究により，パターンに基づく方法と機械学習を組み合わせることで，より性能高く変遷情報を取り出せることがわかった．

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，人手により変遷情報の分類を行った．更に，分類を自動で行いたいと考え，機械学習により変遷情報の分類を行った．その結果，学習データの事例数の多い分類では F 値が 6 割以上であった．

目 次

表 目 次

図 目 次

第1章　はじめに

物事の変遷を知ることは，その物事の知識を会得する際に重要なことである．変遷を知るためには一般的にWebや検索エンジン，または書籍を使用して情報を得る方法があげられるが，これらの方法では人手では網羅的に収集することが困難であり，かつ多大な労力を要する．変遷を知ることを自動で簡単に行うことができれば非常に便利である．

堀らは[1]その手始めとして，論文のタイトル，著者名のデータを使用し，分野の変遷関係，人物の変遷関係(いわゆる先生と弟子のような関係)を自動で推定した．次に，Fanら[2]は法則の変遷情報をWikipediaから抽出した．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した．

これらの研究は，問題点として学術分野間，師弟間，法則間という限定された変遷情報の種類についての抽出であったことが挙げられる．そこで，本研究では，より多くの種類の変遷情報を自動で，より高性能で取得することを大きな目的とする．

この目的を達成するため，本論文では，以下の研究を行う．

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．(4章)

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．(5.1節)

3. 機械学習を利用して変遷情報の自動分類も行う．(5.2節)

本研究の主張点を以下に整理する．

- 大規模テキストから変遷情報を取り出すという特色のある研究対象を扱った．
- パターンで変遷情報を含む可能性のある個所を抜き出し，そこから機械学習でより高性能に変遷情報を抜き出す手法を提案した．この手法はパターンを用いるだけの

手法よりも性能が高いことを確認した．本研究の実験において提案手法は 0.9 という高い F 値を得た．ただし，ここでの F 値はパターンに基づく方法で抽出できた正しい変遷情報の個数に相当するものを再現率の分母にして算出したものである．

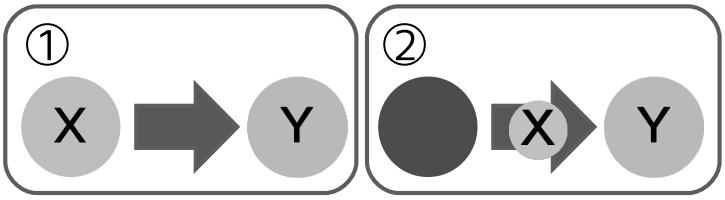
- 変遷情報の人手に基づく分類を行った．これは変遷情報を扱う際の理論的基礎として今後役立つと考える．また，機械学習を使用し，自動的に分類を行った．

第2章　変遷情報

本研究において変遷情報とは，二つの事物 X，Y において，X が時の流れとともに移り変わり Y となった，または，X が影響を及ぼして Y になったという変化のことを変遷情報とする．そして，X → Y と表記し，X，Y のペアを変遷情報と呼ぶ．

＜変遷情報とは＞

X，Yにおいて，
①Xが時の流れとともに移り変わりYとなった変化
②Xが影響を及ぼしてYになったという変化



2

図 2.1: 変遷情報

変遷情報の例

^^e2^^91^^a0 の例 1：イクラ (派生元 X)　→　サケ (派生先 Y)
^^e2^^91^^a0 の例 2：琉球 (派生元 X)　→　沖縄 (派生先 Y)
^^e2^^91^^a0 の例 3：合成樹脂 (派生元 X)　→　ペットボトル (派生先 Y)

^^e2^^91^^a1 の例 1：情報学 (派生元 X)　→　情報工学 (派生先 Y)
^^e2^^91^^a1 の例 2：情報学 (派生元 X)　→　インターネット (派生先 Y)
^^e2^^91^^a1 の例 3：主流文化 (派生元 X)　→　サブカルチャー (派生先 Y)

第3章　先行研究

3.1　研究者と研究分野の変遷情報の抽出

堀ら [1] 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)」のような派性関係の対) を自動的に抽出する方法を提案した．

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．分野名においても同様のことが言える．この仮説に基づいた人名の変遷情報の推定方法を，以下に示す．

手順 1 論文から著者名データ (本論文では著者名と共著の人名を合わせたものを著者名データとする) を抽出し，その中から指定した人名を抽出し人名 A とする．

手順 2 人名 A を含む著者名データを取り出し，その中より (最初の時期によく共起した情報を取り出したいため) 出現年の早いものから 10 件の著者名データを取り出す．

手順 3 その 10 件の著者名データから共起している人名すべてを取り出し，人名 B_i(i は整数．B_i は共起している人名の異なり数だけ設定) とする．出現年の早い順に重みを付け，出現した論文の分だけ人名 B_i ごとにその重みを加算する．

手順 4 B_i のうち，初出現年が人名 A の初出現年よりも早く，重みが最も大きい人名 (人名 B) を人名 A のルーツとする．

また，分野名の変遷情報の推定方法を以下に示す．

手順 1 「言選」[12] を使用し，論文データのタイトル (またはアブストラクトも含めてもよい．ただし本研究ではタイトルのみを利用する．) から名詞連続を取り出し，不要な語を人手で省く．その中から指定した名詞連続を抽出し分野名 A とする．

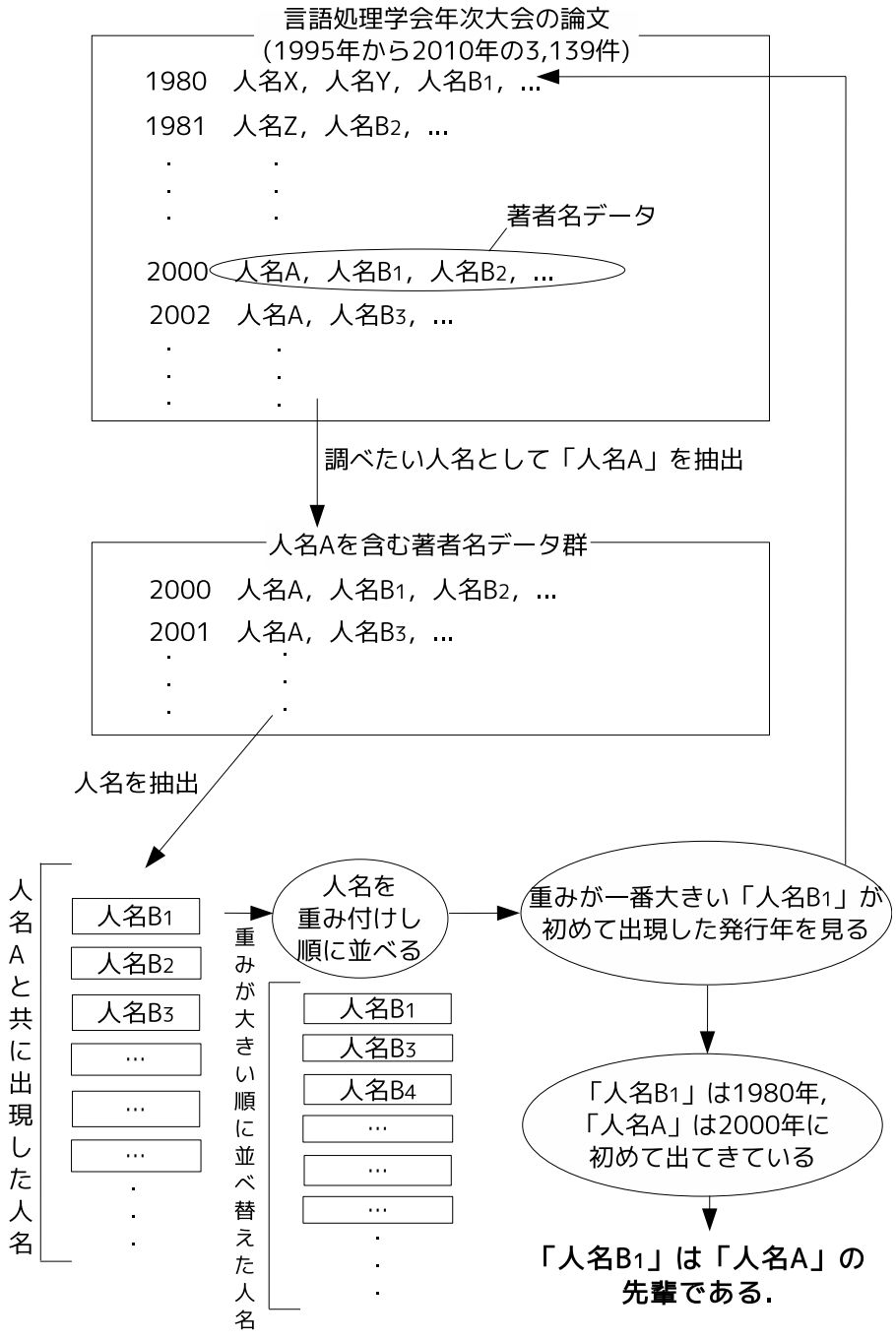


図 3.1: 先行研究の流れの概要図 (人名の変遷情報抽出)

以下，[手順 2] から人名の変遷情報の推定方法と同様.

図 3.1 を用いて説明を行う．図 3.1 は人名のルーツを抽出する例である．まず，調べたい人名が「人名 A」であった場合，全データ (言語処理学会年次大会の論文 1995 年から 2010 年の 3,139 件) から「人名 A」を含む著者名データを抽出する．次に,「人名 A」と共に出現した人名を抽出し，出現年，出現回数により重みを付け，その重みが一番大きいもの (ここでは「人名 B_1」) を抽出する．ここで,「人名 B_1」が初めて出現した年を見て,「人名 A」よりも早かった場合,「人名 B_1」は「人名 A」のルーツとなる．

この方法により，人名と分野名においての変遷情報を抽出することができた．

しかし，この方法では学術分野間，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．

3.2 法則の変遷情報の抽出

Fan ら [2] は Wikipedia から，ヒューリスティックルールおよび機械学習を用いて法則の変遷情報を抽出した．

まず，法則の発見年を予測するため，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.

次に，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(基本法則) と法則 B(関係法則) 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，変遷情報を抽出した．また，教師あり機械学習により，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判断する方法を提案した．

この方法で，変遷情報を抽出した．しかし，この方法では法則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．

3.3 ALAGINの「意味的関係抽出サービス」

ALAGINの「意味的関係抽出サービス」[13]は，パターンを入力すると，大量の文書からパターンに適合した文を抽出することができる，Stijnらの研究[14]を元にしたサービスである．このサービスを用いて，4.1節の自動で変遷情報を抽出する方法を実現した．

以下に詳細な説明を載せる．この説明は,「意味的関係抽出サービスマニュアル」[15]を参考にしている．このサービスでは，統計的な手法を用いた半自動処理により，約6億のWeb文書から効率的に大量の単語対，例えば「原因-結果」「トラブル-予防策」「食材-効能」などの種々の意味的関係を持つ単語対を作成することができる．

「原因-結果」関係インスタンスの例

(「-」の左の単語が「原因」で，右の単語が「結果」)
連鎖球菌 - 化膿性関節炎,EBウイルス - 伝染性単核球症,
ツボカビ - カエルツボカビ症, 断層 - 直下型地震, 煤塵 - 環境問題,
フロン - 地球温暖化問題, トラウマ - PTSD,
ヒューマンエラー- 重大事故, 過冷却 - 結露, 窒素肥料 - 地下水汚染

「トラブル-予防策」関係インスタンスの例

(「-」の左の単語が「トラブル」で，右の単語が「予防策」)
情報漏えい - 暗号化ソフトウェア, 不正アクセス - ファイヤーウォール機能,
床ずれ - エアマット, 鳥害 - 防鳥ネット, 手荒れ - ラノリン,
老化 - ガラクタン, 壁内結露 - 羊毛断熱材, 尿モレ - 立体ギャザー,
白とび - NDフィルター, 腐食 - クロームメッキ

「食材-効能」関係インスタンスの例

(「-」の左の単語が「食材」で，右の単語が「効能」)
にんにく- 精力増強，ウコン - 吐き気,
酢 - 疲労回復，ハトムギ - むくみ,
お茶 - 口臭予防, プーアル茶 - 消化, ざくろ - 美容, ゴーヤ - バテ防止,
ペパーミント - クールダウン，クランベリー - 抗酸化作用

このサービスでは「Xから派生するY」などのパターン(シードパターン)を入力する

と，シードパターンと同様な意味関係を持つ類似したパターンを自動で作成し，シードパターンと自動で作成した類似パターンに合致した X，Y を Web 文書から自動的に抽出する．

ただし，特定の意味的関係に絞ったとしても，その知識は様々な言語パターンで書かれているため，大量のインスタンスを獲得するには大量の言語パターンが必要という問題がある．それらを人手で用意する作業は非常に高コストである．

このサービスは人手コストを最小限にするため，少数の言語パターン (以降，シードパターンと呼ぶ) を入力するだけで稼働するように設計されている．その鍵は，シードパターンと同じ意味的関係を表す，一種の言い換えとなる言語パターン (以降，類似パターンと呼ぶ) を自動学習する機能にある．類似パターンの学習は，同じインスタンスを獲得できるパターン同士は良い言い換えであるという考えに基づいている．例えば，シードパターンとして「X が Y の原因になる」「Y の原因である X」を入力すると，これらと同じインスタンスを獲得しやすい「X によって起こる Y」「X で Y が発生」「Y を招く X」など，多くの人がすぐには思いつきにくい言語パターンも含め，大量の類似パターンを学習してくれる．最終的には，学習された全類似パターンを用いて大量のインスタンスを獲得する．

さらに，このサービスは，単語の意味的なカテゴリの情報 (以降クラスと呼ぶ) を用い，曖昧な言語パターンをうまく活用できるよう工夫している．曖昧な言語パターンとは，複数の異なる意味的関係を表せるもので，例えば「X による Y」という言語パターンは「ノロウイルスによる食中毒」ならば因果関係,「X 社による製品 Y」ならば会社と製品の関係といった具合に様々な関係を表す．曖昧な言語パターンは，X，Y に当てはまる単語の意味に制限を付けることで，その曖昧性を解消することができる．単語のクラスを [クラス名] と書くことにすると，例えば「X による Y」という言語パターンは,「[生物] による [症状]」ならば因果関係,「[組織] による [製品]」ならば会社と製品の関係となる．このように単語のクラスの対毎に異なる言語パターンと考えることで曖昧性を解消できる．すると，例えば「X が Y の原因になる」など因果関係を表す言語パターンの言い換えとしては,「[生物] による [症状]」など因果関係を表す意味カテゴリのペアを持つ言語パターンが学習されるようになる．実際にはこれらのクラスは「生物」「症状」などの意味的なラベルで表されているのではなく，同じような意味を持つと自動判定された単語に同じ ID(1 から 500 までの数字) が割り当てられたものとなる．このサービスの基本的なデータ量は以下の通りである．

基本的なデータ量

抽出対象の文書数:約 6 億ウェブページ
対象とする単語数:約 100 万
単語クラス数:500
クラス対最大数:250,000 (=500 × 500)
利用可能な言語パターン:約 58,700,000 種類

このサービスによって，膨大なデータから意味的関係のインスタンスを獲得することができる．

また，このサービスでは取得時に詳細なオプションも設定することができる．

3.4 その他の関連研究

その他の関連研究としては以下のものがある．

川中ら [4]，[5] は，ソーシャルブックマークにおける概念を記述するタグを相互情報量に基づく方法により解析し，概念の派性関係の対 (例えば「SNS(ルーツ) → mixi(派生概念)」のような対) を自動的に抽出した．

松尾ら [6] は Web 上の情報を用いて共起の強さから人物の関係性の強さを推定し，かつ「共著関係」や「同研究室関係」などの社会的関係性を判別し，その情報が示された人間関係ネットワークを作成した．

Referral Web[7] は，ある人物から対象人物への繋がりを Web 上の情報から順次発見していくものである．ある人物の名前と共起する名前を抽出し，更にその名前から次の名前を抽出するという方法を用いている．

原田ら [8] はある単語で検索した Web ページ集合から固有表現抽出により人物名を抽出し，独自に定義した共起度を使用し，共起関係から人物の関係を表すネットワークを抽出する方法を提案している．

村田ら [9] は検索エンジンで検索された件数を使用して Web ページ間の関係を発見する手法を提案している．

Adar ら [10] はブログ上での情報の流れについて，テキストの類似度，リンク，時間の情報から解析するモデルを提案した．

丹羽ら [11] はソーシャルブックマークにおけるユーザベースの共起度とドキュメントベースの共起度を比較し，Synonym と呼ばれる同じ意味で用いられる語を共起度の高い精度で発見する手法を提案した．

第4章　変遷情報の抽出

まず，多くの種類の変遷情報を自動で取り出したいという課題に対し，変遷情報を自動で抽出する方法を提案する．

4.1 節では，パターンに基づく変遷情報の抽出を行う．

4.2 節では，機械学習に基づく変遷情報の抽出を行う．

4.1　パターンに基づく変遷情報の抽出

大量の文から人手で作成したパターンを利用して，変遷情報を自動で抽出するという方法を提案する．この手法はパターンを用いて行っているため，以後「パターンベース法」と呼ぶ．

4.1.1 節では，パターンベース法の詳細な説明を行う．

4.1.2 節では，パターンベース法によって行った文の抽出実験の結果を記載する．

また，4.1.3 節では，パターンベース法によって抽出した文が判定を含むか否かの判定を行う．

4.1.1　提案手法

図 4.1 にパターンベース法の概要図を示す．大量の Web 文書から，図 4.1 のように「X を元にした Y」「X から生まれた Y」「X から派生した」などの変遷情報が抽出できると思われるパターンを適用する．パターンを適用し自動抽出するために，本研究では，ALAGIN の「意味的関係抽出サービス」を利用する．このサービスにより，図 4.1 右下のように，パターンが適合した文が自動的に抽出することができる．

また，このサービスでは取得時に詳細なオプションも設定することができる．オプションの 1 つに，シードパターンから作成した類似パターンを利用できるというものがある．

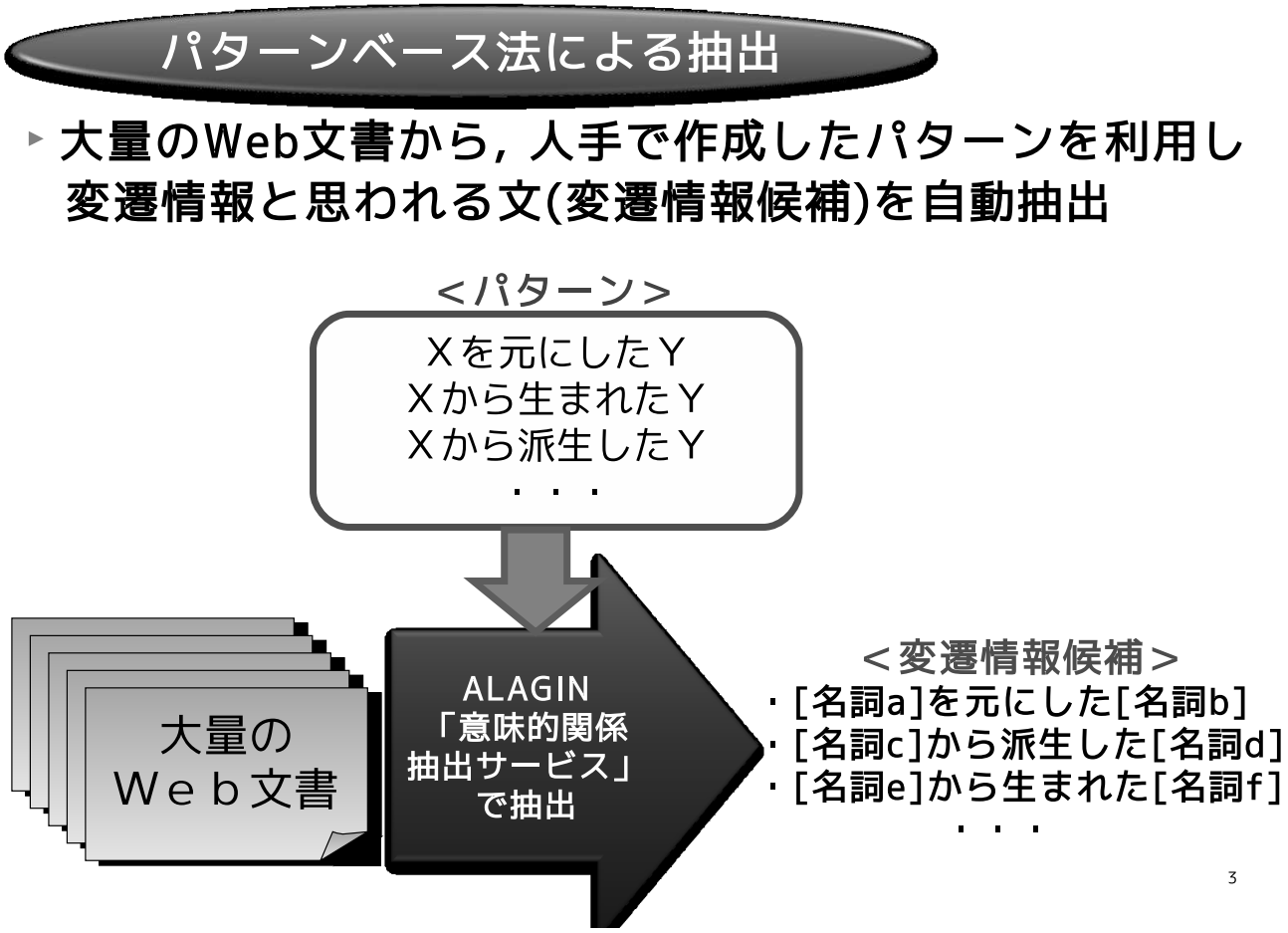


図 4.1: パターンベース法

本研究での事前の実験では，このオプションを利用すると，抽出されたものの性能は低かった．このため，本研究では自動で作成される類似パターンは利用せず，入力したパターンのみを利用する．

4.1.2 実験結果

大量の文から人手で作成したパターンを利用して，変遷情報を自動で抽出した．シードパターンは34個入力した．以下に図4.1の＜パターン＞に対応している，使用するパターンを載せる．

使用するパターン

1, <X から産まれた Y>
2, <X から生まれた Y>
3, <X から生まれる Y>
4, <X から生み出された Y>
5, <X から生み出される Y>
6, <X から誕生した Y>
7, <X から派生した Y>
8, <X から派生する Y>
9, <X で生まれた Y>
10, <X によって生まれた Y>
11, <X を元にした Y>
12, <X を語源とする Y>
13, <Y から生まれた X>
14, <Y の引き金となる X>
15, <Y の基となる X>
16, <Y の元となった X>
17, <Y の元となる X>
18, <Y の元になった X>
19, <Y の元になっている X>
20, <Y の元になる X>
21, <Y の元凶である X>
22, <Y の原因物質である X>
23, <Y の原料である X>
24, <Y の原料になる X>
25, <Y の材料である X>
26, <Y の材料となる X>
27, <Y の成分である X>
28, <Y の素である X>
29, <Y の素となる X>
30, <Y の素になる X>
31, <Y は X から生まれ>
32, <Y は X から生まれる>
32, <Y は X から生まれる>
33, <Y は X から派生>
34, <Y を発生させる X>

また，以下に抽出できた文の例を載せる．

抽出できた文の例 1

しかしこのヒドロキシルラジカルの元となる過酸化水素を除去する酵素も人間は持っています。

抽出できた文の例 2

氷温の状態で素材をねかせると、細胞が凍るまいとして旨みの元になる成分を作り出し、熟成が進みます。

抽出できた文の例 3

鹿児島県肉用牛改良研究所は２４日，体細胞クローン技術で生まれたウシの細胞から作った「再クローン牛」が生まれたと発表した。

この実験で抽出できた文は 3,115 文である．しかし，3,000 文程度では少ないと思われる．これはシードパターンが少なかったこと, シードパターンの質が悪かったことが原因として挙げられる．解決方法としてはシードパターンの見直しが考えられる．また，サービスの類似パターンを利用することも考えられる．

4.1.3 抽出した文の判定

抽出した文は変遷情報を含んでいるか含んでいないかの判定を行う．

変遷情報の含み方に基づく分類

4.1 節でパターンベース法で抽出した文の判定は以下の typeA〜F の分類に基づいて行う．なお，本論文においては,「X から派生する Y」などのパターンベース法でのシードパターンで X と Y に当てはまる名詞により，X と Y が変遷関係にあるかを判断する．

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．

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 の名詞から変遷情報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，変遷情報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．

次節より，各 type ごとの詳しい説明と例を挙げる．例の下線部がシードパターンに当てはまっているものであり，二重線部の単語ですが X と Y にあたる単語である．

1.type-A について

type-A は，X，Y が明らかに変遷情報であり，知見の得られる事例である．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-A の例 1

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。

「神経 (Y) の元になる幹細胞 (X)」は明らかに変遷情報であり，知見の得られる事例である．

type-A の例 2

エビやカニなど甲殻類や昆虫の 外皮 (X)の成分であるキチン (Y) もセルロースと並ぶ生体構造ポリマーです。

「外皮 (Y) の成分であるキチン (X)]」のみでは何の「外皮」なのかわからないが，Y を修飾している「昆虫の」によって詳しく説明されているため，はっきりと知見の得られる情報となる．また，並列関係となっている「エビやカニなど甲殻類や」でより詳しい情報が得られる．

type-A の例 3

肌の保湿と細胞の活性化に大きく貢献するグリシン、プロリン、くすみ (Y)の元になるアンモニア (X) を代謝させるアルギニン、美白作用のあるシステインをつくり出すメチオニン、ヒト成長ホルモンを活性化させるグルタミン酸など、アミノ酸がバランスよく組み合わされています。

「くすみ (Y) の元になるアンモニア (X)」は明らかに変遷情報であり，知見の得られる事例である．また，補足ではあるが,「システインをつくり出すメチオニン」という他の変遷情報も得られる事例である．

2.type-B について

type-B は，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である．

ここでの文の構造は，広い意味を持つ名詞が名詞述語文の述語に相当するものであり名詞述語文の主語の部分がその名詞の具体的な内容を示している構造 (「として」,「とされる」などの補語を取る文も名詞述語文と同様に扱う)，X，Y に対して離れた個所で修飾関係にある表現が広い意味を持つ名詞の内容を示している構造とする．

type-B の例 1

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。

「お酒の雑味 (Y) の元となる成分 (X)」は,「成分」が一般的に広い意味を持つ名詞であり，知見が得にくいと考えられる．しかし,「成分」は「タンパク質や粗脂肪などの」によって修飾されているため,「タンパク質や粗脂肪などの成分」となるため，知見が得られると考える．

type-B の例 2

このＮＡＧはヒアルロン酸の構成成分であり、うるおいの (Y)素となる成分 (X) とされています。

「うるおい (Y) の素となる成分 (X)」は,「成分」が一般的に広い意味を持つ名詞であり，知見が得にくいと考えられる．しかし,「成分」は名詞述語文の述語に相当するものであり名詞述語文の主語の部分「ＮＡＧは」がその名詞の具体的な内容を示しているため，知見が得られると考える．

type-B の例 3

ＲＤＦは家庭や事業者から排出された可燃性のごみを押し固めてつくられる燃料で、電気 (Y)を発生させる熱源 (X) として利用することができます。

「電気 (Y) を発生させる熱源 (X)」は,「熱源」が一般的に広い意味を持つ名詞であり，

知見が得にくいと考えられる．この文は「として」という補語を取る文であり,「ＲＤＦは」が「熱源」の具体的な内容を示しているため，知見が得られると考える．

3.type-C について

type-c は，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 の名詞から変遷情報として知見の得られる事例である．

type-C の例 1

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。

「原因菌」は一般的に広い意味を持つ名詞であり，あまり知見が得られないが,「にきび (Y) の元となる原因菌 (X)」のように X，Y の両方の名詞が得られれば知見が得られる．

type-C の例 2

汚れ (Y)の元となる有機成分 (X) を分解.

「有機成分」は一般的に広い意味を持つ名詞であり，あまり知見が得られないが,「汚れ (Y) の元となる有機成分 (X)」のように X，Y の両方の名詞が得られれば知見が得られる．

type-C の例 3

そうするとそこから 痒み (Y)の元になる抗原 (X) が入り込んでさらに痒くなる、という悪循環を繰り返す皮膚の病気です。

「抗原」は一般的に広い意味を持つ名詞であり，あまり知見が得られないが,「痒み (Y) の元になる抗原 (X)」のように X，Y の両方の名詞が得られれば知見が得られる．

4.type-D について

type-D は，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，変遷情報として知見の得られない事例である．

type-D の例 1

Ｊ Ａ Ａ Ａ ３ ０ 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

「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-D の例 2

真珠貝は、自分の体から貝殻を作るために 貝殻 (Y)の元になる物質 (X) を出す、そこに核が入っていると、その核にも貝殻の元になる物質が付いて真珠となる。

「物質」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-D の例 3

解体材 (Y)から生まれる商品 (X) をしばらく追って行こうと思っています。

「商品」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．

5.type-E について

type-E は，単に場所を指定しているため，変遷情報として知見の得られない事例である．

type-E の例 1

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。

「自社畑」は場所を示している．

type-E の例 2

日本 (X)で生まれた伝統武道の空手道 (Y) は、今や世界１７３ヵ国に普及し、４千万人もの愛好者を持つポピュラーなスポーツとなりました。

「日本」は場所を示している．

type-E の例 3

地中に張り巡らされた木の根は、土や石をおさえ、山崩れを防ぎ、また、森 (X)で生まれたミネラル (Y) たっぷりのおいしい水は、植物や川魚、野生動物たちの命の源です。

「森」は場所を示している．

6.type-F について

type-F は，単に状態を表しているため，変遷情報として知見の得られない事例である．

type-F の例 1

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。

「同じような境遇 (X) で生まれた」は，単に「組織の先輩 (Y)」の状態を表している．

type-F の例 2

アルコール中毒患者 (X)から生まれた子供 (Y) の混濁角膜の免疫組織化学的検討

「アルコール中毒患者 (X) から生まれた」は，単に「子供」の状態を表している．

type-F の例 3

未熟児で生まれた幼児に対する教育相談活動の取組

「未熟児 (X) で生まれた」は，単に「幼児」の状態を表している．

分類結果

4.1 節のパターンベース法により抽出した 3,115 文からランダムで 100 文取りだし，4.1.3 節の typeA〜F に分類した．

結果を表 4.1 に載せる．

表 4.1: 変遷情報の含み方に基づく分類の結果

分類	type-A	type-B	type-C	type-D	type-E	type-F
個数	65/100	6/100	5 /100	17/100	5/100	2/100

100 文中の 6 割以上が type-A に分類された．

抽出した文の判定

本研究では，type-A〜C と分類された文が変遷情報を含む文とし，それ以外は変遷情報を含まない文と仮定した．

表 4.2: 抽出した文の判定

	文数 (文)
変遷情報を含む文	76
変遷情報を含んでいない文	24

この結果より，パターンベース法で抽出した文には 76 %の変遷情報が含まれていることがわかった．

4.2　機械学習に基づく変遷情報の抽出

4.1.3 節では，76 %の適合率で変遷情報が抽出できることがわかった．本研究では，この精度をより向上させるため，パターンベース法に機械学習法を追加するという方法を提案する．パターンベース法で取得したものを，教師あり機械学習で変遷情報を含むか否か判定し，機械学習が変遷情報と判断した文を変遷情報として取り出す．

4.2.1 節では，機械学習に基づく変遷情報の抽出方法の詳細な説明を行う．

4.2.2 節では，機械学習に基づく変遷情報の抽出を行う．

4.2.1　提案手法

図 4.2 に提案手法の概要図を載せる．図 4.2 中央上のようにパターンベース法で判定した文を学習データとして機械学習を行う．大量の文から，機械学習を使用し，図 4.2 右下のような変遷情報を含む文を抽出する．

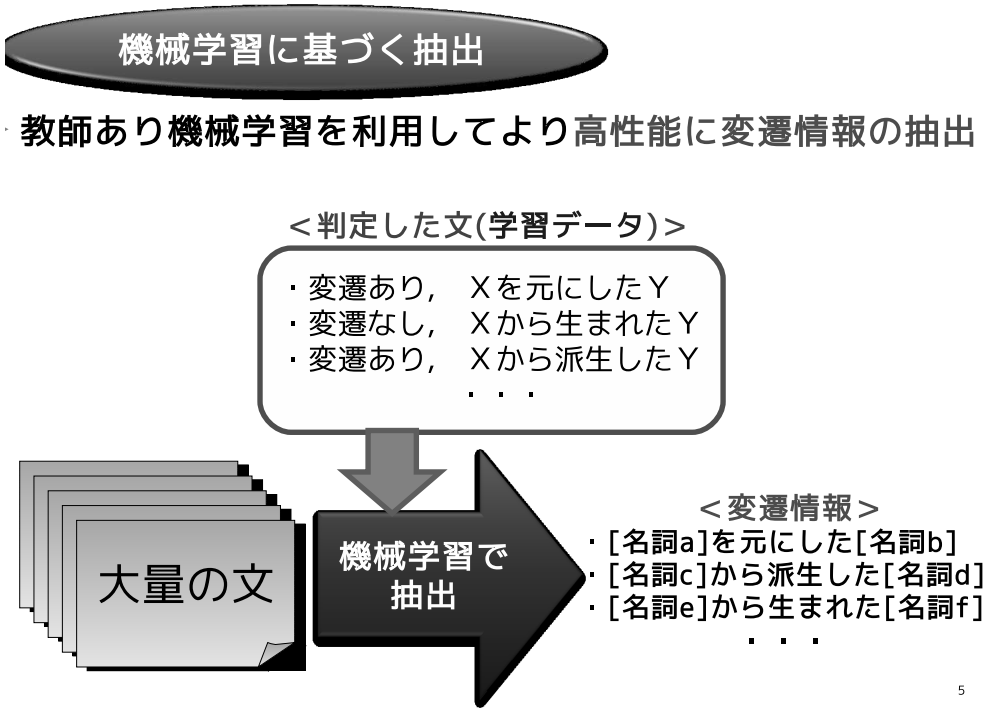


図 4.2: 提案手法の概要

機械学習にはサポートベクトルマシン法，最大エントロピー法を使用する．次節に手

法の説明を載せる．この説明は，村田らの論文 [16] を参考にしている．また，機械学習に使用した素性も記載する．

サポートベクトルマシン法 (SVM)

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間隔 (マージン) が大きいもの (図 4.3 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 [17, 18].

$$
\begin{aligned}
f(\mathbf{x}) &= sgn\left(\sum_{i=1}^{l} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right) \\
b &= -\frac{max_{i,y_i=-1} b_i + min_{i,y_i=1} b_i}{2} \\
b_i &= \sum_{j=1}^{l} \alpha_j y_j K(\mathbf{x}_j, \mathbf{x}_i)
\end{aligned}
\tag{4.1}
$$

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，

$$
sgn(x) = \begin{cases} 1 & (x \geq 0) \\ -1 & (otherwise) \end{cases} \tag{4.2}
$$

であり，また，各 α_i は式 (4.4) と式 (4.5) の制約のもと式 (4.3) の $L(\alpha)$ を最大にする場合のものである．

$$
L(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(\mathbf{x_i}, \mathbf{x_j}) \tag{4.3}
$$

$$
0 \leq \alpha_i \leq C \ (i = 1, ..., l) \tag{4.4}
$$

[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．

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{4.5}$$

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．

$$K(\mathbf{x}, \mathbf{y}) \quad = (\mathbf{x} \cdot \mathbf{y} + 1)^d \tag{4.6}$$

C, d は実験的に設定される定数である．本論文では C,d はともにすべての実験を通して 1 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 (4.1) 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．

サポートベクトルマシン法は分類の数が 2 個のデータを扱うもので，通常これにペアワイズ手法を組み合わせて用いることで，分類の数が 3 個以上のデータを扱うことになる [19]．

ペアワイズ手法とは，N 個の分類を持つデータの場合，異なる二つの分類先のあらゆるペア ($N(N-1)/2$ 個) を作り，各ペアごとにどちらがよいかを 2 値分類器 (ここではサポートベクトルマシン法[2]) で求め，最終的に $N(N-1)/2$ 個の 2 値分類器の分類先の多数決により，分類先を求める方法である．

本論文のサポートベクトルマシン法は，上記のようにサポートベクトルマシン法とペアワイズ手法を組み合わせることによって実現される．

[2]本論文の 2 値分類器としてのサポートベクトルマシンは，工藤氏が作成した TinySVM[18] を利用している．

最大エントロピー法 (ME)

最大エントロピー法は，あらかじめ設定しておいた素性 $f_j(1 \leq j \leq k)$ の集合を F とするとき，式 (4.7) を満足しながらエントロピーを意味する式 (4.8) を最大にするときの確率分布 $p(a,b)$ を求め，その確率分布にしたがって求まる各分類の確率のうち，もっとも大きい確率値を持つ分類を求める分類とする方法である [20].

$$\sum_{a\in A, b\in B} p(a,b)g_j(a,b) = \sum_{a\in A, b\in B} \tilde{p}(a,b)g_j(a,b) \quad for\ \forall f_j\ (1 \leq j \leq k) \tag{4.7}$$

$$H(p) = -\sum_{a\in A, b\in B} p(a,b)\ log\,(p(a,b)) \tag{4.8}$$

ただし，A, B は分類と文脈の集合を意味し，$g_j(a,b)$ は 文脈 b に素性 f_j があってなおかつ分類が a の場合 1 となりそれ以外で 0 となる関数を意味する．また，$\tilde{p}(a,b)$ は，既知データでの (a,b) の出現の割合を意味する．

式 (4.7) は確率 p と出力と素性の組の出現を意味する関数 g をかけることで出力と素性の組の頻度の期待値を求めることになっており，右辺の既知データにおける期待値と，左辺の求める確率分布に基づいて計算される期待値が等しいことを制約として，エントロピー最大化 (確率分布の平滑化) を行って，出力と文脈の確率分布を求めるものとなっている．