# 多面的な用語説明を生成するためのテキスト分類手法

筑波大学図書館情報専門学群
三條場旭彦
筑波大学大学院図書館情報メディア研究科
藤井敦

発表者：伊藤輝将

# はじめに

・科学技術や文化の急速な発展によって、様々な用語について調べる機会が増えている

↓

こうした調べものを支援するツールに辞典や検索エンジンがある

| | 辞典 | 検索エンジン |
|---|---|---|
| 長所 | ・人手で情報が統制されており質が高い | ・情報量が多く新語に対応できる |
| 短所 | ・未収録語が調べられないため情報量が少ない | ・検索された情報は統制されていないため質が低い |

# はじめに(2)

↓そこで、、、

- 辞典と検索エンジン両者の長所を統合することで、Web上の雑多な情報から説明情報を抽出し、様々な用語に対する説明を辞典のように構築する研究を行っている
- 本研究は、ある用語に関する多面的な説明情報を生成するために、Web上の雑多な情報を「説明の観点」に基づいて分類する手法を提案する

# 用語説明の観点

- 動物名では「分布」、「形態」など「病名」では「診断」や「治療」などと用語の種類によって観点は異なる

↓そこで、、、

-人手で作成されたフリー百科事典Wikipediaから用語の種類ごとに観点を抽出し、用語説明のモデルを生成する

-さらに生成したモデルを使用して、Web上の雑多な情報を観点に分類する

# 用語説明の分類手法

- 概要

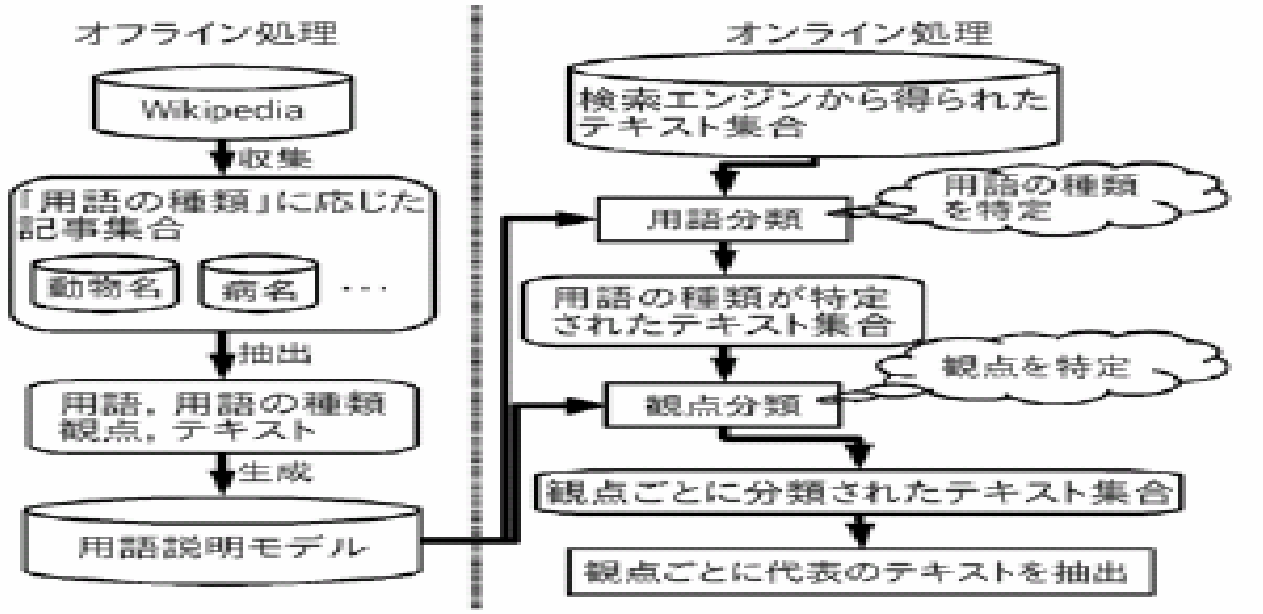


図 1: 本研究で提案するテキスト分類手法

# 用語説明の分類手法(2)

- 用語の種類によって説明に必要な観点が異なるため、「用語分類」と「観点分類」の二段階で分類を実行する
  -どちらの分類もサポートベクターマシンを使用し、One vs Rest法を用いて多値分類に拡張した
- 用語分類
  →SVMのスコアが高い用語の種類にテキストを分類する
- 観点分類
  →分類された用語の種類に対応する観点候補のいずれかに分類する
  例:「動物名」→「分布」、「形態」など
  「スポーツ名」→「ルール」、「用具」など

# 用語の種類に応じた記事の収集

- 記事の収集はWikipediaの記事に付与されている「カテゴリ」を使用する

- Wikipediaに付与されているカテゴリの単位と、一般的な用語の種類が必ず一致するとは限らない

  →収集する用語の種類に対応するカテゴリを目視で特定した

# 用語の種類に応じた記事の収集（2）

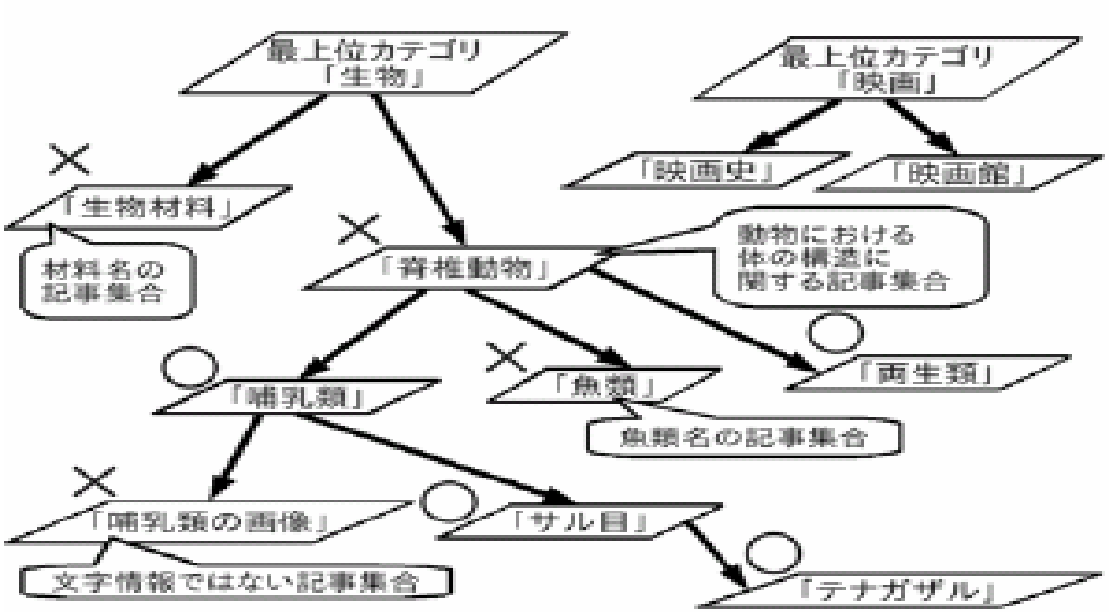


図 2: Wikipedia カテゴリの階層

# 用語説明のモデルの作成

- 用語に応じた観点を設定するため、Wikipediaの「セクション」を利用する
  -収集した記事から50件以上の用語で使用されているセクションを観点とした

  (50件を超える説明情報が得られない用語の種類は、使用頻度が高い3件のセクションを観点とした)
- 以下のセクションは説明の観点として不適切であると判断し、削除した

  概要、概説、概論、備考、脚注、出典、その他、外部リンク、関連、関連文献、関連図書、関連書籍、関連事項、関連記事、関連項目、関連リンク、関連カテゴリ、参考、参考書籍、参考文献、画像

# 用語説明のモデルの作成(2)

- ChaSenを使用し、観点ごとの説明情報を形態素解析する
  - -観点分類の素性

    名詞、動詞、形容詞、形容動詞、アルファベット、未知語
  - -用語分類の素性

    上記6つ、用語に含まれる単語(用語自体も形態素解析する)
- 素性の値は全て1とした

# テキスト分類の実行例

- 「ハクビシン」をYahoo!で検索して得られた上位100件のスニペットを分類した

- 括弧内の数字は左から順に、「該当する観点に正しく分類された件数」、「該当する観点に分類された件数」、「Yahoo!検索における順位」であり、太文字が観点に対応する記述である

# テキスト分類の実行例(2)

- 生態（4/5，74 位）

  **ハクビシンは雑食性であり、特に果実類を好む**ことから秩父地域の観光果樹園のぶどうを ... ハクビシンは、食肉目ジャコウネコ科で東南アジア. を中心とした亜熱帯から熱帯地域に生息している。 ... これは、ハクビシン. が持つ ...

- 分布（6/55，4 位）

  白鼻心】、 masked palm civet、[学名：Paguma larvata] 哺乳（ほにゅう）綱食肉目ジャコウネコ科の動物。**中国南部、チベット、インド、マレー半島、ボルネオ島、スマトラ島、 ...**

- 形態（5/7，16 位）

  ハクビシン・アライグマ・アナグマ・イタチ・イノシシ 害獣でお困りならお任せください! ... ハクビシン ジャコウネコ科 **体長/50〜70cm 体重/3〜5kg** ... 年々増え続けていますが、ハクビシンやアライグマなどは、 ...

- 人間との関係（15/23，22 位）

  ハクビシン. 飼育園館. 多摩動物公園・井の頭自然文化園. 生息地. 東南アジア、中国など。 **日本では四国、静岡など。最近は東京でも見られることがある。** ... ハクビシンがむかしから日本にいたのか、それとも帰化動物なのか、 ...

- 特徴（2/2，37 位）

  ハクビシンも実は国外から持ち込まれた外来種です。 ... ハクビシンはその名前の由来のとおり、**鼻筋にとおる白いくっきりとした線と、長くしなやかなしっぽが特徴的です。 ... 足の指の数を比べてみると、タヌキは 4 本ですが、ハクビシンでは 5 本あります。 ...**

- 分類（4/5，1 位）

  トップ **せきつい動物亜門 哺乳類 (哺乳綱) 食肉目 裂脚亜目 ジャコウネコ科 ハクビシン. ハクビシン (C)Kojo TANAKA. 分類. 哺乳類 食肉目裂脚亜目 ジャコウネコ科 ...** インドのカシミール地方、中国、台湾、マレー半島、 ...

- 歴史（1/3，84 位）

  **ハクビシンは、明治維新以前に既に国内に持ち込まれていたと言われています。 江戸期の鳥獣戯画にハクビシン ... 本市では、ハクビシンが平成１４年８月、環境省による移入種対応方針に移入種として掲載されたことから、以来移入種として対応しています。...**

# 評価実験

- 本研究の目的は、Web上の雑多なテキストを分類すること

  ↓しかし

  今回は「整った」テキストであるWikipediaの記事を分類し、その正解率を評価した

- 収集した記事をセクションごとに分割したテキストを1件のデータとした

- 5分割の交差検定によって用語分類と観点分類の正解率を計算し、結果を以下の表に示す

# 評価実験(2)

表 1: 用語の種類における記事数と観点数と分類の正解率

| 用語の種類 | 記事数 | 観点数 | 用語分類 | | 観点分類 | |
|---|---|---|---|---|---|---|
| | | | 正解記事数 | 正解率 | 正解記事数 | 正解率 |
| 動物名 | 1317 | 7 | 1201 | 91.19 | 1199 | 91.04 |
| 映画名 | 1169 | 5 | 1154 | 98.72 | 938 | 80.24 |
| 病名 | 878 | 8 | 853 | 97.15 | 674 | 76.77 |
| 企業名 | 618 | 3 | 556 | 89.97 | 602 | 97.41 |
| 人名 | 500 | 4 | 454 | 90.80 | 302 | 60.40 |
| 植物名 | 276 | 3 | 212 | 76.81 | 240 | 86.96 |
| 数学用語 | 228 | 3 | 198 | 86.84 | 140 | 61.40 |
| 虫名 | 203 | 3 | 140 | 68.97 | 171 | 84.24 |
| 化学用語 | 201 | 3 | 156 | 77.61 | 169 | 84.08 |
| 料理名 | 190 | 3 | 137 | 72.11 | 146 | 76.84 |
| 情報工学用語 | 103 | 3 | 61 | 59.22 | 61 | 59.22 |
| 魚類名 | 96 | 3 | 44 | 45.83 | 62 | 64.58 |
| スポーツ名 | 86 | 3 | 66 | 76.74 | 65 | 75.58 |
| 法学用語 | 56 | 3 | 40 | 71.43 | 30 | 53.57 |
| 建築学用語 | 55 | 3 | 17 | 30.91 | 30 | 54.55 |
| 電気工学用語 | 49 | 3 | 7 | 14.29 | 18 | 36.73 |
| 天文学用語 | 38 | 3 | 21 | 55.26 | 20 | 52.63 |
| 獣医学用語 | 34 | 3 | 0 | 0 | 15 | 44.12 |
| 地質学用語 | 24 | 3 | 0 | 0 | 7 | 29.17 |
| 物性物理学用語 | 23 | 3 | 0 | 0 | 5 | 21.74 |
| 平均 | 307.2 | 3.6 | 265.85 | 86.54 | 244.7 | 79.66 |

表 2: 「企業名」に関する観点分類の内訳

| 観点 | 記事数 | 分類された観点 | | | 正解率 |
|---|---|---|---|---|---|
| | | 沿革 | 事業所 | 主な商品 | |
| 沿革 | 469 | 437 | 0 | 2 | 99.54% |
| 事業所 | 116 | 2 | 112 | 2 | 96.55% |
| 主な商品 | 63 | 8 | 2 | 53 | 84.13% |

表 3: 「人名」に関する観点分類の内訳

| 観点 | 記事数 | 分類された観点 | | | | 正解率 |
|---|---|---|---|---|---|---|
| | | 経歴 | 略歴 | 著書 | 人物 | |
| 経歴 | 219 | 142 | 48 | 28 | 1 | 64.84% |
| 略歴 | 133 | 69 | 55 | 7 | 2 | 41.35% |
| 著書 | 83 | 0 | 0 | 83 | 0 | 100% |
| 人物 | 65 | 20 | 9 | 23 | 13 | 20% |

# おわりに

- Wikipediaから用語説明モデルを生成し、web上のテキストを説明の観点に基づいて分類する手法を提案した

- 現在：テキストを分割せずに分類している
  →複数の観点が混在し、重複する場合がある
- 今後：文などの単位で分割して分類することで、冗長性の少ない説明を生成する必要がある