

Wikipedia マイニングによる百科事典的サーチ

藤井 敦[†] 三條場旭彦[†]

[†] 筑波大学 〒305-8550 茨城県つくば市春日 1-2

あらまし 本研究は、多種多様の用語について多面的な説明を効率よく調べるための百科事典的なサーチ手法を提案する。説明に必要な観点は、「動物名」や「病名」といった用語の種類によって根本的に異なるため、人手に依存した手法では大規模化が困難である。本研究は、Wikipedia をマイニングすることで、用語の種類ごとに観点の構造に関するテンプレートを構築する。さらに、Wikipedia の記事集合を用いて、与えられたテキストがどのような観点に基づく説明であるのかを判定するための分類器を学習する。その結果、ある用語について検索エンジンで収集した Web ページやスニペットを観点に分類し、その用語に関する多面的な説明情報を生成することが可能になった。

キーワード 情報検索、百科事典、用語説明、Wikipedia、マイニング

Encyclopedic Search Powered by Wikipedia Mining

Atsushi FUJII[†] and Akihiko SANJOUBA[†]

[†] University of Tsukuba Kasuga 1-2, Tsukuba, 305-8550 Japan

Abstract This paper proposes an encyclopedic search method that enables users to consult various terms from multiple viewpoints. Because viewpoints required for explanation are different depending on the type of a term, such as animals and diseases, it is difficult to manually produce a large scale system. We mine Wikipedia for a prototype of a viewpoint structure for each term type. We also use articles in Wikipedia for a machine learning method, which categorizes a given text into an appropriate viewpoint. As a result, given texts retrieved by a search engine in response to a term, such Web pages and snippets, our method categorizes those texts based on viewpoints and produces an expository text for the term in question.

Key words Information retrieval, Encyclopedias, Term descriptions, Wikipedia, Mining

1. 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公私を問わず増えている。そのため、World Wide Web 上の様々なツールを使うことが多い。代表的なツールには、Google や Yahoo!などの「検索エンジン」と Wikipedia などの人手で編集された「事典」がある。両者には、情報の量と質という点において、それぞれ長所と短所がある。

検索エンジンは、億単位のページ集合が検索の対象であり、提供される情報の量が多いという利点がある。しかし、検索される情報が体系化されておらず、必要のない情報も含まれるため、情報の質が低い。

他方において、事典は、説明を目的とした情報に限定され、かつ項目等によって情報が統制されているため、情報の質が高いという利点がある。しかし、人手による編集に依存しているため、調べたい言葉が必ず登録されているとは限らない。また、説明の内容が著者の視点に偏るという問題もある。すなわち、情報の量において問題がある。

本研究は、検索エンジンと事典の長所を統合して、有用性が高い調べ物のツールを実現することを目的とする。具体的には、Wikipedia をマイニングすることによって「用語説明が編集される仕組み」を解明し、用語説明に関するモデル（用語説明モデル）を構築する。さらに、そのモデルに基づいて検索エンジンの結果を組織化することで「百科事典的なサーチ」を実現する。

用語説明モデルの構築において、動物名や病名といった対象によって説明に必要な観点が異なる点に着目した。例えば「動物名」ならば「生態」や「形態」、「病名」ならば「症状」や「治療」といった観点に基づいて説明される。そこで、Wikipedia から用語の種類に応じて異なる観点の構造を学習し、さらに観点ごとに固有の単語分布を学習する。その結果、例えば「ハクピシン」に関する Web 検索の結果に含まれる複数のページやスニペットから「生態」、「形態」、「分布」などの観点に対応するテキストを抽出し「ハクピシン」に関する事典的な情報を組織化することが可能となる。

2. 本研究の位置付け

2.1 Cyclone の要約手法との比較

筆者らが開発した CYCLONE^(注1)の要約機能[3]は、説明の「観点」に着目し、ある用語について説明している複数のテキストから、観点ごとに代表文を抽出する。観点の設定と観点ごとに代表文を抽出する規則の作成は人手で行っている。現在は、情報処理用語を対象とし「定義」「例示」「目的」など12の観点をを用いている。しかし、用語の種類によって観点が異なる点を考慮していない。また、観点ごとの代表文を抽出する規則を人手で作成している。その結果、現状の CYCLONE では、情報処理用語の要約に限定されている。

2.2 Wikipedia を用いた要約手法との比較

Biadsky ら[1]は、人物情報の要約に Wikipedia を利用した。Wikipedia から人物に関する記事を収集し、人物の説明に使われる文のモデルを学習した。この手法は、Wikipedia を用いて要約に使用する情報を選別する点では本研究と類似する。しかし、Biadsky らは説明の観点を使用していない点が本研究と異なる。また、Biadsky らが人物情報の要約に限定していることに対して、本研究は用語の種類を制限せず、見出し語の種類に応じて異なる観点を集合を使い分ける点も異なる。

2.3 複数文書要約手法との比較

本研究の目的は、複数文書要約[4]とも関連する。複数文書要約に関する既存の研究において、観点に基づく研究の事例は少ない。サービスに関する評判情報の要約[2]では、例えば、ホテルに関する評判情報から「立地」や「価格」などの観点に基づいて要約が生成される。しかし、観点や観点の名称は評判情報の内容から抽出される。それに対して、本研究は、Wikipedia を外部情報として利用することによって、要約対象の情報からは得ることができない情報を補間する点異なる。同様に、本研究は、Clusty^(注2)のように Web 検索結果の内容を分析してクラスタリングする手法とも異なる。

3. 百科事典的サーチの手法

3.1 概要

本研究で提案する百科事典的サーチの概要を図1に示す。図1は、事前に行う「用語説明モデルの構築」と検索結果のテキストが与えられた段階で実行する「検索結果の組織化」に大別される。検索結果の組織化では、「りんご病」のような用語を検索質問として Web を検索した結果から、複数のテキスト（ページまたはスニペット）を収集し、用語説明モデルに基づいて観点ごとに分類する。さらに、観点ごとに代表的なテキストをユーザに提示する。

3.2 用語説明モデルの構築

本研究で構築する用語説明モデルとは、「病名」や「動物名」といった用語の種類に応じて、説明の観点を列挙したテンプレートである。さらに、ある用語について検索されたテキスト

が与えられると、用語の種類を特定し、その見出し語に対応する観点の候補から適切な観点を特定するための分類器である。

まず「人名」「動物名」「病名」といった用語の種類ごとに、Wikipedia の記事から観点構造のテンプレートを作成する。ここでは、Wikipedia の記事にある「目次」に着目する。図2は「破傷風」に関する Wikipedia の記事である。図2の「目次」には「1. 原因」「2. 症状」「3. 治療」などの項目が並んでいる。本研究では、1つの項目を1つの観点として使用する。

しかし、あらゆる病名の記事で目次の項目が完全に一致するわけではない。そこで、ある用語の種類（例えば「病名」）ごとに複数の記事を収集し、使用頻度が高い目次項目を観点として選択する。観点の選択基準として、使用頻度に関する閾値を設定する方法や、使用する観点数の上限を決めておく方法があり、目的等に応じて適宜使い分けが必要がある。

さらに、用語や観点を分類するために、該当する Wikipedia の記事集合から複数の分類器を学習する。ただし、分類の目的によって使用する学習データが異なる。用語の種類を分類するためには、用語の種類ごとに記事集合をまとめて、1つのカテゴリに対応する学習データを作る。他方で、観点を分類するためには、観点ごと記事集合をまとめて、1つのカテゴリに対応する学習データを作る。

本研究では、機械学習にサポートベクターマシン (SVM) を使用し、さらに多値分類に拡張した。多値分類への拡張方法、素性、値にはそれぞれ複数の選択肢がある。予備実験の結果、One-Vs-Rest 法を使って多値分類に拡張し、単語ユニグラムを素性として、値は全て定数とした場合に分類精度が最も良かったため、これらを採用する。

3.3 検索結果の組織化

検索結果の組織化では、「りんご病」などの用語について検索されたテキストを入力し「用語分類」と「観点分類」を順番に実行する。用語分類では、用語に関する分類器を用いて「りんご病」の説明テキストを「人名」「動物名」「病名」のいずれかに分類する。観点分類では、分類された用語の種類に応じて観点を分類する。図1の例では「病名」と分類された場合、説明テキストを病名に対応する「原因」「症状」「治療」のいずれかに分類する。SVM に基づく One-Vs-Rest 法では、分類結果ごとにスコアが計算されるため、観点ごとにスコアが高いテキストを抽出し、ユーザに提示する。

なお、予備実験[5]では、用語分類と観点分類について良好な分類精度が得られた。実験の規模を拡張しながら、さらなる評価を行う予定である。

3.4 実行例

Yahoo! で「ハクピシン」を入力して検索した上位100件のスニペットを分類した。本研究で作成した用語説明モデルには、「動物名」に対して「生態」や「分布」など合計7種類の観点が対応していた。以下、観点ごとに分類された代表スニペットを示す。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スニペットの件数」「該当する観点に自動分類されたスニペットの件数」「Yahoo! 検索における当該スニペットの順位」である。太文字の箇所が観点に対応する記述である。ま

(注1): <http://cyclone.slis.tsukuba.ac.jp/>

(注2): <http://clusty.jp/>

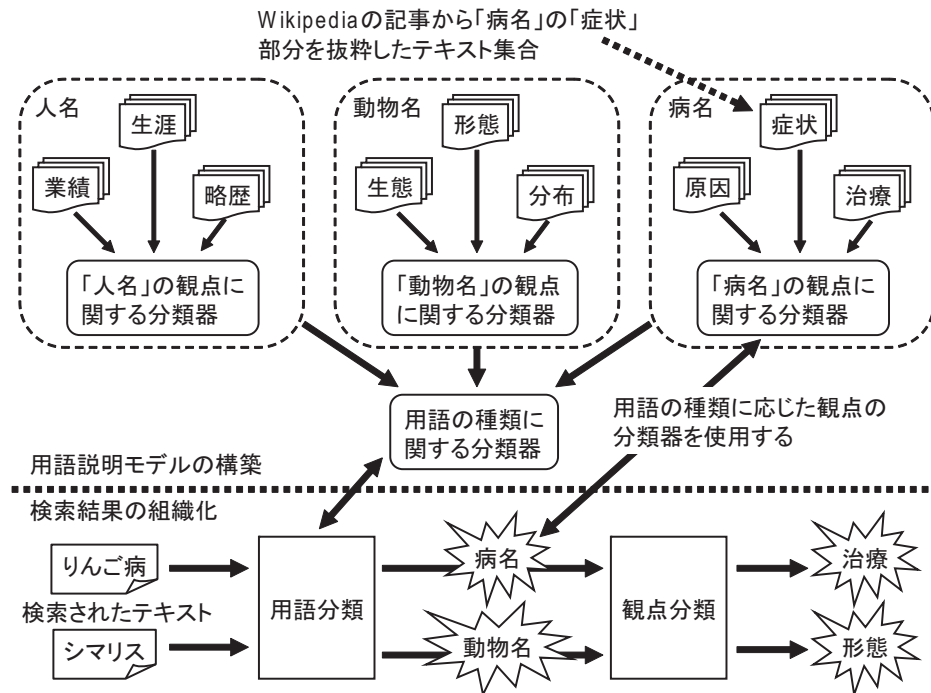


図 1 百科事典的サーチの概要



図 2 「破傷風」に関する Wikipedia の記事

た、スニペットを抽出したページの URL を示す。

- 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 ...

<http://www.pref.saitama.lg.jp/A06/BQ23/02/siken/H13/h13.06-13.pdf>

- 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、 ...

<http://100.yahoo.co.jp/detail/ハクビシン/>

- 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、...

<http://www.naturelink.jp/>

- 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、...

[http://www.tokyo-zoo.net/encyclopedia/
species_detail?code=343](http://www.tokyo-zoo.net/encyclopedia/species_detail?code=343)

- 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、鼻筋にとある白いくっ
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。...

[http://www.nat.pref.ibaraki.jp/kyoiku/kids/
yagai/yagai03.html](http://www.nat.pref.ibaraki.jp/kyoiku/kids/yagai/yagai03.html)

- 分類 (4/5, 6 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、...

[http://contents.kids.yahoo.co.jp/zukan/
mammals/card/0168.html](http://contents.kids.yahoo.co.jp/zukan/mammals/card/0168.html)

- 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。...

[http://www.city.kamakura.kanagawa.jp/
bikaiese/higai_hakubishin2.html](http://www.city.kamakura.kanagawa.jp/bikaiese/higai_hakubishin2.html)

この実行例では、本手法で分類された 7 件のスニペットを読め
ば、「ハクビシン」について多面的な説明を得ることができた。
それに対して、Yahoo! 検索の結果を上位から読んだ場合には、
上記 7 つの観点に関する情報を得るために、33 件のスニペット
を読む必要があった。

括弧内の数字を見ると、「分布」に分類されたスニペットの
多くは誤りであった。これは、ハクビシンの目撃情報などに地
名がかかっていると、「分布」に分類されやすいことが原因で
あった。

本稿執筆当時、Wikipedia における「ハクビシン」の記事に
は、「形態」「歴史」「分布」「分類」に関するセクションはな
い。しかし、本手法は動物名に関する記事の集合から代表的な
セクションを観点として抽出するため、Wikipedia において個

別の記事では欠落している観点を補うことができた。

分類したスニペットには、上記 7 種類以外の観点として「ハ
クビシン」をテーマにした「著作」や「名称の由来」があった。
これらの用語説明モデルに含まれない観点は、検索されたスニ
ペットの内容を分析して抽出するしかない。今後は、カテゴリ
分類とクラスタリングの併用について検討する必要がある。

比較のために、Clusty で「ハクビシン」を検索し、クラスタ
リングした結果「Sars」のような関連語のクラスタが生成され
た一方で、「日記」や「クビに」のような無関係なクラスタも生
成された。これは、本手法の方が事典的な検索に有利であるこ
とを示す例である。

多義語の例として「キーウィ」で検索されたスニペットを分
類した。用語分類において、鳥のキーウィについて記述された
スニペットは「動物名」に分類され、果物のキーウィについて
記述されたスニペットは「植物名」に分類された。さらに、こ
れらのスニペットは観点分類によって動物名や植物名に固有の
観点に分類された。本手法によって多義や多観点を考慮した検
索を実現できることが分かった。

4. おわりに

本研究の目的は、検索エンジンと事典の長所を統合して大規
模な情報を統制する点にある。Wikipedia から説明の観点に関
する構造を抽出し、種々のテキストを観点に基づいて分類する
問題に帰着させた。本手法の特長は、与えられたテキストがど
のような用語について書かれているかを特定し、その結果に基
づいて分類すべき観点の候補を変更する点にある。Wikipedia
のように協調的な人手編集による事典は今後も発展するだろう。
しかし、爆発的に増える情報を自動的に統制する技術も必要で
ある。情報の統制において自動化が困難な事象を特定し、人手
編集との棲み分けや共存について検討する必要がある。

謝 辞

本研究の一部は、文部科学省科研費特定領域研究「情報爆発
時代に向けた新しい IT 基盤技術の研究」(課題番号: 19024007)
によって実施された。

文 献

- [1] Fadi Biadisy, Julia Hirschberg, and Elena Filatova. An unsupervised approach to biography production using Wikipedia. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pp. 807–815, 2008.
- [2] Sasha Blair-Goldensohn, Ryan McDonald, George Reis, Tyler Neylon, Kerry Hannan, and Jeff Reynar. Building a sentiment summarizer for local service reviews. In *WWW2008 Workshop on NLP Challenges in the Information Explosion Era*, 2008.
- [3] Atsushi Fujii and Tetsuya Ishikawa. Summarizing encyclopedic term descriptions on the Web. In *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 645–651, 2004.
- [4] Inderjeet Mani. *Automatic Summarization*, chapter 7, pp. 169–208. John Benjamins, 2001.
- [5] 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, 2009.