Webと特許情報を事典的に活用するシステムの高度化

東京工業大学大学院情報理工学研究科准教授 **藤井　敦**

PROFILE 1998年東京工業大学大学院博士課程修了。博士（工学）。筑波大学大学院准教授等を経て、2009年より現職。自然言語処理、情報検索、Webマイニング、特許情報処理の研究に従事。

1 はじめに

知的財産権の一つである特許権は、高度な発明の保護を目的としている。日本では年間約34万件の特許が出願され、多様な専門分野に関する発明が蓄積されている。特許に内在する人間の英知を体系化し、活用することができれば、今日の高度情報化社会において産業上の価値が高い。

特許には発明に関する新語や専門用語が多く含まれている。筆者は、World Wide Webと特許情報から種々の用語に関する説明情報を抽出し、さらに複数の説明情報を組織化することで、百科事典的なコンテンツを自動構築する研究を行っている［1,2,3］。また、構築したコンテンツに対して、見出し語、関連語、同義語、質問文、関連語マップといった多様な手段によって検索する機能を開発している。当該研究成果は、事典検索システムCyclone［4］として公開中である。

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。

事典は、説明を目的とした情報に限定され、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、調べたい用語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。

本研究の目的は、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である。本研究の特長は、既存の事典であるWikipediaを分析することによって「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）を構築する点にある。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し、「事典的な検索」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」には「生態」や「形態」といった観点が使われ、「病名」には「症状」や「治療」といった観点が使われる。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、動物の「ハクビシン」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、「ハクビシン」に関する事典的な情報を組織化することを可能とする。

ここで、「Wikipediaの存在が前提であれば、Wikipediaの記事を読めば十分ではないか？」という疑問が生じるかもしれない。この問いに対する答えは「No」であり、本研究には二つの意義がある。まず、Wikipediaの未登録語についてWikipediaと同じ

ような観点の構造で説明を得ることができる。さらに、Wikipedia の登録語に対しても、不足している観点を補うことや、一般の Web ページや特許情報から幅広く説明を収集することができる。

本稿では、筆者らの先行研究［1,2］について解説し、さらに用語説明モデルの構築を自動化する新たな手法［3］について解説する。

2 事典的検索の手法

2.1 概要

本研究で提案する事典的検索の概要を図 1 に示す。図 1 は、事前に行う「用語説明モデルの構築（上部）」と検索結果のテキスト集合が与えられた段階で実行する「検索結果の組織化（下部）」に大別される。用語説明モデルは Wikipedia の記事集合を用いて構築する。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個々のテキストを分類する。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。

2.2 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、「動物名」や「病名」といった用語の種類に応じて、説明の観点を列挙したプロトタイプである。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、その用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。

具体的には、「人名」、「動物名」、「病名」といった用語の種類ごとに Wikipedia の記事から観点の構造を抽出する。ここでは、Wikipedia の記事にある「目次」に着目する。例えば、「破傷風」に関する Wikipedia 記事の目次には、「原因」、「症状」、「治療」などの項目（セクション）が並んでいる。本研究では、一つのセクションを一つの観点として使用する。

しかし、あらゆる病名の記事でセクションが完全に一致するわけではない。そこで、用語の種類（例えば「病名」）ごとに複数の記事を収集し、使用頻度が高いセクション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る必要がある。

さらに、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一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一つのカテゴリに対応する学習データを作る。分類器の学習にはサポートベクターマシン（SVM）を使用し、さらに One-Vs-Rest 法を使って多値分類に拡張する。

2.3 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。図 1 の例で説明すると、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」に分類されたため、病名に対応する「原因」、「症状」、「治療」のいずれかに説明テキストを分類する。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。

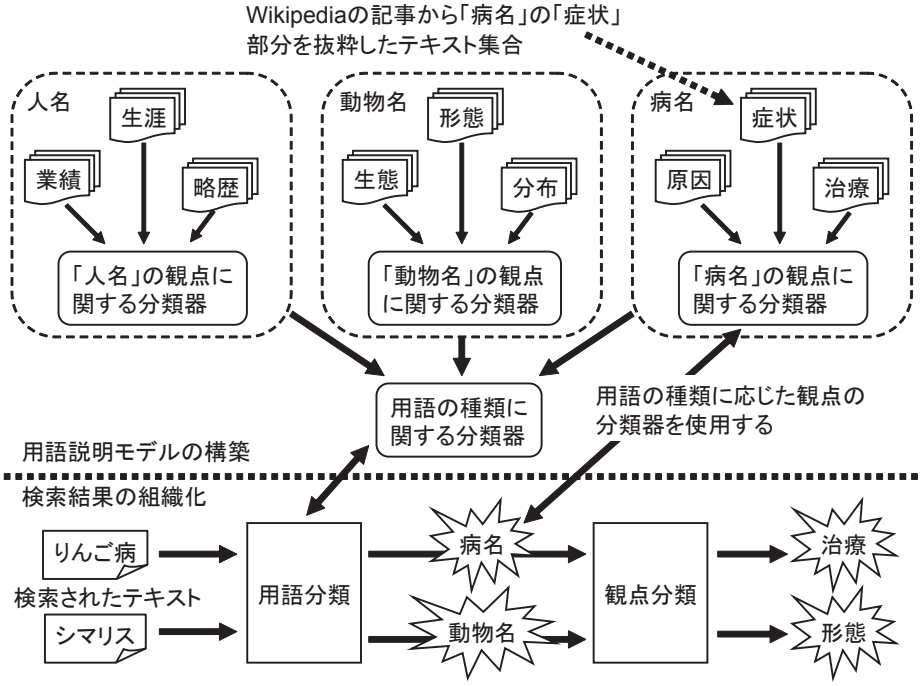

図 1　用語説明のモデル化による事典的検索の概要

3 用語説明モデル構築の自動化

2.2 で説明した用語説明モデルの構築では、「人名」、「動物名」、「病名」といった用語の種類ごとに Wikipedia の記事を収集する必要がある。ここで問題となるのは、用語の種類をどのように特定するのかという点である。Wikipedia 記事には一つ以上のカテゴリが付与されているものの、「動物」のカテゴリには「脊椎動物」や「家畜」などに混ざって、「動物を題材とした作品」のように「動物」の体系に適さない記事も含まれている。そこで、筆者らの先行研究［1,2］では Wikipedia 記事の収集は人手で行っていた。それに対して、最新の研究［3］では、Wikipedia の記事集合をクラスタリングすることで、用語の種類に相当する「用語クラスタ」を自動的に特定する手法を提案した。

用語クラスタの自動特定について、約 5000 件の Wikipedia 記事を対象として実験を行った。各記事には、以下に示す 10 カテゴリのいずれか一つが人手によって事前に付与されている。

動物、映画、病気、企業、人物、植物、虫、料理、魚類、スポーツ

表 1 は、Wikipedia 記事のクラスタリングによって自動的に特定された用語クラスタと、各用語クラスタに対応付けされた観点の一覧である。表 1 の「用語クラスタ」には、各クラスタにおいて代表的なカテゴリを示している。「用語クラスタ」の欄を見ると、「動物」が「動物 1」と「動物 2」に分割されており、その代わりに「魚類」がなくなっている。しかし、「観点」の欄を見ると、用語クラスタごとに人間の直感に合う観点の名称が抽出されていることが分かる。ただし、今回の実験では日常語に関する記事を対象とした。今後は技術系の専門用語を対象とした実験も必要である。

表１　自動構築された用語クラスタと観点の対応

用語クラスタ	観点
虫	特徴、分類、生態、形態
スポーツ	歴史、ルール
映画	キャスト、スタッフ、ストーリー、あらすじ、登場人物
病気	治療、症状、原因、検査、診断、分類、疫学、予後、病態、歴史、予防
動物１	特徴、歴史
料理	歴史、作り方
企業	沿革、事業所、主な商品、関連会社、歴史、会社概要、主な製品
人物	経歴、略歴、人物、来歴・人物、著書、来歴、生涯、生い立ち、パーソナル、エピソード
植物	特徴、利用、分類、主な種
動物２	生態、形態、人間との関係、分類、分布、特徴、亜種、利用、近縁種

4 おわりに

本研究は、Wikipediaという既存の事典を用いて、用語説明が編集される仕組みをモデル化した。さらに、構築した用語説明モデルを用いて、検索エンジンで得られたテキスト集合を説明の観点に基づいて分類し、事典的な検索を可能にした。本研究の特長は、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、科学研究費補助金基盤研究（B）（課題番号 22300050）によって実施された。

参考文献

[1] 藤井敦．特許情報を専門用語辞典として活用するシステム，Japio 2008 Year Book, pp.196-199, 2008.

[2] 藤井 敦. Webと特許情報を事典的に活用するシステム. Japio 2009 Year Book, pp.172-177, 2009.

[3] Atsushi Fujii, Yuya Fujii, and Takenobu Tokunaga. Effects of Document Clustering in Modeling Wikipedia-style Term Descriptions. Proceedings of the 8th International Conference on Language Resources and Evaluation, pp. 2543-2546, 2012.

[4] http://cyclone.cl.cs.titech.ac.jp/