特集論文

Wikipediaマイニング

近未来チャレンジキックオフ編

Wikipedia Mining

Challenge for Realizing Early Profits, The Kick Off

中山 浩太郎 Kotaro NAKAYAMA
東京大学知の構造化センター
The Center for Knowledge Structuring, The University of Tokyo
nakayama@cks.u-tokyo.ac.jp, http://www.cks.u-tokyo.ac.jp/nakayama

伊藤 雅弘 Masahiro ITO
大阪大学大学院情報科学研究科
Dept. of Multimedia Eng., Graduate School of Information Science and Technology, Osaka University
ito.masahiro@ist.osaka-u.ac.jp, http://www-nishio.ist.osaka-u.ac.jp/~ito

Erdmann, Maike Maike ERDMANN（同　上）
erdmann.maike@ist.osaka-u.ac.jp, http://www-nishio.ist.osaka-u.ac.jp/~pasture

白川 真澄 Masumi SHIRAKAWA（同　上）
shirakawa.masumi@ist.osaka-u.ac.jp, http://www-nishio.ist.osaka-u.ac.jp/~shirakawa

道下 智之 Tomoyuki MICHISHITA（同　上）
michishita.tomoyuki@ist.osaka-u.ac.jp, http://www-nishio.ist.osaka-u.ac.jp/~michishita

原 隆浩 Takahiro HARA（同　上）
hara@ist.osaka-u.ac.jp, http://www-nishio.ist.osaka-u.ac.jp/~hara

西尾 章治郎 Shojiro NISHIO（同　上）
nishio@ist.osaka-u.ac.jp, http://www-nishio.ist.osaka-u.ac.jp/~nishio

keywords: Wikipedia Mining, Social Media, Ontology, Thesaurus

Summary

Wikipedia, a collaborative Wiki-based encyclopedia, has become a huge phenomenon among Internet users. It covers a huge number of concepts of various fields such as arts, geography, history, science, sports and games. As a corpus for knowledge extraction, Wikipedia's impressive characteristics are not limited to the scale, but also include the dense link structure, URL based word sense disambiguation, and brief anchor texts. Because of these characteristics, Wikipedia has become a promising corpus and a new frontier for research. In the past few years, a considerable number of researches have been conducted in various areas such as semantic relatedness measurement, bilingual dictionary construction, and ontology construction. Extracting machine understandable knowledge from Wikipedia to enhance the intelligence on computational systems is the main goal of "Wikipedia Mining," a project on CREP (Challenge for Realizing Early Profits) in JSAI. In this paper, we take a comprehensive, panoramic view of Wikipedia Mining research and the current status of our challenge. After that, we will discuss about the future vision of this challenge.

1. は じ め に

Wikipedia は，インターネットを通じて誰でも編集可能なオンライン百科事典であり，ここ数年で爆発的に成長したソーシャルメディアの一種である．Wikipedia は，Web ブラウザを通じて自由に編集可能なことから，ユーザ同士が迅速かつ容易にコンテンツを編集するための情報基盤を提供してきた．この結果，多くのユーザが精査することによって質の高いコンテンツの量は増え，さらなるユーザを獲得するというサイクルを形成している．

インターネット上に新しい情報共有の基盤を作り上げた Wikipedia は，社会現象としても興味深いが，研究者にとっては魅力的な新しい研究用のリソースへと成長してきた．これを証明するように，ここ数年で Web マイニングを始め，人工知能や自然言語処理，情報検索など幅広い研究分野で研究の基盤リソースとして利用され，その有用性が示されてきた．特に，統計的手法に基づく自然言語処理の研究領域では，ある程度まとまった量の高品質なテキスト情報がコーパスとして必要であったが，Wikipedia はこの要件を満たし，GFDL（GNU Free Documentation License）に基づくコピーレフトなライセンス形態で利用しやすいという点から，標準的な言語リソースの一つになりつつある．

Wikipedia マイニングとは，このように急成長してき

たWikipediaを解析対象とし，有用な知識を抽出する研究の総称である．Wikipediaマイニングに関する研究領域は多岐にわたり，概念同士の関係性を数値化する連想関係抽出，より詳しい関係の種類を抽出するオントロジ構築，さらには語の曖昧性解消アルゴリズムの基盤情報としての利用や情報検索への応用などに関して研究が進められてきた．特に，意味情報を中心とするウェブの実現を目指す「セマンティックWeb」においては，概念同士の関係を抽出するための基盤リソースとしてWikipediaに注目が集まっている．

このような状況の下，筆者らはWikipediaを解析することで，人工知能研究をはじめとする幅広い分野で利用可能な基盤リソースを構築することを目指し，近未来チャレンジ「Wikipediaマイニング」を提案した．本チャレンジでは，Wikipediaを解析することにより，有用な知識を抽出し，幅広いアプリケーションで利用可能な汎用的な基盤リソースを構築していくことを目的とする．特に,「連想関係抽出（Relatedness Measurement）」「関係抽出（Relation Extraction）」「対訳関係抽出の高度化（Translation Enhancement）」の三つの研究を大きな柱として進める．これは，これら三つの研究がWikipediaマイニングにおいて最も活発な分野であることもあるが，意味情報を中心としたWWWを実現するために必要な要素だからである．そして，５年以内にこれらのリソースを使った情報検索やテキスト分類などの実アプリケーションを開発・公開し，その有用性を示すことを目指す．

本稿では，Wikipediaの解析に関する背景を解説した後，関連する研究を踏まえながら本チャレンジの現状を解説する．そして，最後に今後の達成目標，学術的・社会的意義を議論する．

2. Wikipediaの特徴

「Wikipedia」は，Wiki[Leuf 01]をベースにした大規模Web百科事典である．Wikiをベースにしているため，誰でもWebブラウザを通じて記事内容を変更できることが大きな特徴である．この編集の容易さがインターネットユーザの書き込みを促進し，今では一般的な概念だけでなく，文化，歴史，数学，科学，社会，テクノロジなどの幅広い分野をカバーし，普遍的な概念から新しい概念に至るまで，非常に膨大なコンテンツが網羅されている．その記事数は既に180万（2007年6月英語のみカウント）を超えており，世界最大の百科事典であるBritannicaの記事数が，全60巻で約65,000記事であることと比較した場合，実に30倍近い数の記事が網羅されていることになる．

Wikipediaは，この幅広いトピックの網羅性以外にも研究の対象として見たときに，URLによる概念識別，250を超える言語のサポート，半構造化データなど興味深い特徴をいくつか持つ [中山 07]．詳しい特徴については[中山 07]に委ねるが，以降，本節ではWikipediaのコーパスとしての特徴について特に重要な部分を略解する．

2·1 URLによる概念識別 (Disambiguation)

URLにより語彙の一意性が確立されている点は，Wikipediaの大きな特徴の一つである．通常の自然言語処理の精度を低下させる要因の一つに語の曖昧性回避問題がある．通常のコーパスでは，曖昧性の高い単語に対してコンテキストを解析しながらその曖昧性を解消するのが一般的だが，曖昧性回避の精度がその後の処理に影響を与えるため，ある程度の精度で解析できたとしても最終的な解析の精度を下げてしまう要因になっていた．

一方，Wikipediaでは一つのURL（ページ）に一つの概念が割り当てられており，多義性がURLによって解決されている点が特徴である．例えば,「Apache」は強いコンテキスト依存性を持つ単語であり，先住民族を示す場合もHTTPサーバや軍用ヘリを示す場合もある．Wikipediaでは，これら三つの概念は別のページで管理されており，それぞれ「http://en.wikipedia.org/wiki/Apache」「http://ja.wikipedia.org/wiki/Apache_HTTP_Server」「http://en.wikipedia.org/wiki/AH-64_Apache」という別々のURLが割り当てられている．

このように，概念とURLが一対一で対応していることは，概念の関係を解析する際に多義性やコンテキスト依存性の影響を受けずに解析できることを示している．

2·2 カテゴリリンク

カテゴリリンクは，ある記事（概念）がどのようなカテゴリに属するかを指定するためのリンクである．すべてのカテゴリには専用のページ（カテゴリページ）が用意されており，カテゴリページはさらに別のカテゴリページに属することが可能である．このカテゴリ構造は，一種のタクソノミー（分類辞書）としての役割を有しており，カテゴリを絞り込みながら記事を検索するような機能を実現するために利用されている．Wikipediaが提供しているカテゴリ検索システム「CategoryTree」[Wikimedia Foundation 08]（以降カテゴリツリー）では，カテゴリを検索することや，カテゴリの階層構造をブラウジングすることが可能である．

しかし，一見階層構造に見えるWikipediaのカテゴリ構造は，一種の擬似的な木構造である．一つのカテゴリページは複数のカテゴリページに属することが可能であり，一部にはループなども存在する．そのため，完全な木構造ではなく，ネットワーク構造となっている．Wikipediaの英語版（2008年5月）を調査したところ，約997万のカテゴリリンクが存在していた．

2·3 半構造化データ

Wikipediaにはカテゴリリンクやリダイレクトリンクなど，いくつかの半構造化されたデータが存在する．こ

れらの情報は，比較的容易に解析可能なデータ形式をしている上に，情報量が豊富であるため，分類的関係や同義語関係の抽出などの研究によく利用されている．さらに,「インフォボックス」も解析が容易な半構造化データとして，Wikipedia 研究では頻繁に利用される．インフォボックスは，各記事において，属性情報を記述するためのテンプレートであり，人に関する記事であれば「生年月日」「血液型」「身長」といった属性情報が記載される．また，都市や国に関する記事であれば,「首都」や「隣接する国」といった属性情報が記述される．インフォボックスは，Wiki のテンプレート機能を利用して記述されており，構文が明確に定義されているため，自然言語処理などを利用せずとも情報を抽出できる．そのため，事象とその属性情報を抽出し，意味的トリプルを抽出するために利用されることが多い．

2·4 250 を超える言語サポート

Wikipedia は 2008 年 4 月の段階ですでに 250 を超える言語をサポートしており，各言語においても広い範囲の記事が網羅されている．ところで，Wikipedia では同じ概念を記述した異なる言語のページは，言語間リンク（Interlanguage Link）と呼ばれる対訳関係を示す特殊なリンクによって結ばれる．日本語と英語の記事の間には 21 万件の言語間リンク（2008 年 2 月）が存在しており，和英電子辞書の EDICT が 12 万件であることと比較したときに十分な数の対訳関係が存在することがわかる．一般的に,（電子）辞書は限られた数の専門家によって更新されるため，最新概念に対する網羅性が弱いという問題があるが，Wikipedia は最新の概念や専門用語に対しての網羅性が高いことから従来の辞書の弱点を補完できるのではないかという期待もある．

3. 関 連 研 究

3·1 関連度計算（Relatedness Measurements）

概念の関連度を数値化する研究は，情報検索のクエリ拡張や文書分類など利用用途が広いことから，情報検索や文書分類の分野で広く研究が進められてきた．また，Wikipedia マイニングの研究の中でも，最も歴史が長く，活発に行われている研究分野の一つである．関連度計算とは，任意の概念ペア間の関連度の強さを数値として算出する処理である．例えば,「コンピュータ」と「メモリ」などは比較的強い関連を持つ概念だが,「コンピュータ」と「ジャガイモ」などは一般的に関連度が高いとは言えない．本研究領域では，このような関連の強さを数値化することを目的としている．また，二つの概念からその関連度を計算するだけでなく，一つの概念から関係の深い概念集合を抽出するような研究もこの研究分野に分類される．

また，関連度計算では関連の強い概念ペアが抽出できるため，次のステップとしてそれらの概念が具体的にどのような関係にあるか，という関係抽出の研究の基本情報として利用することが可能である．

この分野の代表的な研究として，Strube らの WikiRelate! [Strube 06] が挙げられる．Strube らは，WordNet に用いられてきた関連度算出の手法が Wikipedia のカテゴリツリーに適用できることを証明し，複数の指標を統合することで精度が向上することを示した．WikiRelate!では，カテゴリツリーの解析手法をさらに三つの手法に分類している．１）カテゴリ構造におけるパスの長さに基づく手法，２）カテゴリ構造における情報の共有度（直近の共通の祖先が持つ子概念が少ないほど関連度が高い）に基づく手法，３）記事の内容（出現単語のヒストグラムなど）の重複の程度に基づく手法がある．

また，テキスト内容を比較する手法として，Gabrilovich らの研究 [Gabrilovich 07] が挙げられる．Gabrilovich らの研究では，単語やテキストの意味を表現するための手法として，Explicit Semantic Analysis（ESA，明示的意味解析）を提案している．

関連度計算に関する研究全体の課題として，関係抽出やオントロジマッピングへの発展が挙げられる．いくつかの研究では，連想関係抽出によって抽出された概念ペアが，具体的にはどのような関係を持っているか（例：is-a, part-of）を抽出することが重要であると述べている．

実際に Strube らは，[Strube 06] の論文の後，関係抽出の論文 [Ponzetto 07] を発表している．また，さらなる精度向上や網羅性の向上も課題ではあるが，関係抽出やオントロジマッピングを目標とした研究の一部分としての位置づけである場合が多い．

3·2 関係抽出（Relation Extraction）

Wikipedia 研究の中で最も活発な研究分野の一つが関係抽出である．関連度計算の研究では，概念間の関係性の強さを連続的な数値で表現するのに対し，関係抽出に関する研究では，概念間の明示的な意味関係を抽出することを目的としている．例えば，二つの概念が与えられた時に，その間の関係の強さを求めるのではなく，is-a 関係なのか part-of 関係なのか，といったような関係のタイプを抽出することを目的としている．

Wikipedia から概念間の関係を抽出する研究の方向性は，Semantic Web の目標である「意味中心の Web」を実現する基盤技術として必要な大規模 Web オントロジを実現する現実的な方法として注目されている．関係抽出の研究は主に，フリーテキストを利用する方法，カテゴリリンクを利用する方法，インフォボックスを利用する方法に分類される．

フリーテキストを解析することで概念間の関係を抽出する研究としては，PORE [Wang 07]，Dat らの研究 [Nguyen 07]，Aron らの研究 [Culotta 06] などが挙げられる．PORE [Wang 07] は，POL（Positive Only Learn-

ing） [Li 03] をベースにした関係抽出のアルゴリズムを提案している．POL は，Espresso [Pantel 06] に代表されるようなブーストラッピング手法の一種であり，少量の教師データをシード（種）として徐々に正解集合を拡張していく手法である．PORE では，この手法を利用しカテゴリやインフォボックスなどの構造化データを抽出しやすい部分から正解集合を作成した後に，フリーテキストに含まれる共起リンク間の関係を抽出する際の教師データとして利用する方法を提案している．また，Dat らの研究 [Nguyen 07] も，同様に Wikipedia の記事を解析し，二つのエントリ間に含まれるテキストを述語として抽出するアプローチである．Dat らの研究では，単にエントリ間のテキストを抽出するだけでなく，構文解析の結果（構文木）に含まれるパターンを発見することで精度向上を図っている．また，照応解析の方法として，頻出代名詞を利用している．フリーテキストの解析における技術的課題はスケーラビリティと精度である．2008 年 5 月の段階での英語版 Wikipedia には，約 13GB のテキストデータが存在し，これらすべてのテキストを構文解析などの自然言語処理にかけるためには大量の計算機リソースを必要とする．そのため，重要な部分だけに対象を絞って解析することなどが必要となる．また，Wikipedia の記事内では省略表現や代名詞が数多く利用されるため，各文に含まれる主題の照応解析が重要な課題となる．

Wikipedia から概念間の関係を抽出する研究でよく利用されるデータの一つのがカテゴリツリーである．Ponzetto らの研究 [Ponzetto 07] では，主にカテゴリ名に対して文字列照合を行うことによって，Wikipedia のカテゴリリンクを，is-a 関係と not-is-a 関係に分類する手法を提案している．例えば，カテゴリ名 British Computer Scientists と Computer Scientists は，解析して得られた語彙の主要部分（lexical head）を共有しているので is-a 関係に分類される．ここでは，Computer Scientists が British Computer Scientists の語彙の主要部分である．一方で，Crime Comics と Crime のように，Crime Comics の語彙の主要部分（ここでは Comics）が，もう一方のカテゴリ名の先頭に出現しない場合は，not-is-a 関係に分類される．一方，YAGO[Suchanek 07, Suchanek 08] は，Wikipedia と WordNet を利用したオントロジである．YAGO では，オントロジー中のクラスを Wikipedia のカテゴリ名から，そのクラスに属するインスタンスを記事タイトルから収集している．クラスを収集するために，すべての Wikipedia カテゴリから，Conceptual Category を識別しそれをクラスとする．

インフォボックスの解析では，フリーテキストを自然言語処理して関係抽出する手法に対して比較的高精度に構造化されたデータを抽出しやすいのが特徴である．これは，前述の PORE [Wang 07] でもインフォボックスから正解集合を作成していることからもわかる．DBPedia [Auer 07] はインフォボックスから構造化データを抽出する研究の代表例であり，抽出した概念関係を RDF 化し，SPARQL での検索インタフェースを提供している．このことにより，複雑なクエリを処理可能なオブジェクト検索などを実現している．また，DBpedia は意味情報に対して URL を付与し，相互にリンクすることを目標とした「Linked Data」の方向性に基づいて，現在 FOAF（Friend Of A Friend）や GeoNames，DBLP などのさまざまな外部データとの連携が行われている．

3·3 対訳抽出（Translation Extraction）

Wikipedia は 250 以上の言語をサポートし，言語間には多数の言語間リンクが存在する．そのため，単に言語間リンクを抽出するだけでも対訳辞書が構築可能であるが，さらに対訳関係の網羅性を向上させる研究や，抽出精度を向上させようという研究が進められている．

本分野では，単に言語リンクから対訳関係を抽出し，アプリケーションに適用するという研究が多い．例えば，Bouma らの研究 [Bouma 06] や Ferrandez らの研究 [Schönhofen 07] では，多言語に対応した質問応答システムを構築するために Wikipedia の言語リンクを解析して得られた対訳関係を利用している．また，Schoenhofen らの研究 [Ferrández 07] では，言語リンクを解析して固有名詞の翻訳や曖昧性解消を行っている．

一方，言語リンクを対訳関係として抽出する研究のほかに，言語リンクで結ばれたページ同士が多くの場合同じ内容を示していることを利用し，対称コーパス（comparable corpus）を構築する試みも行われてきた．Haghighi らは，対称コーパス（comparable corpus）を Wikipedia から抽出する手法を提案し [Haghighi 08]，実験によりその有効性を示している．Adafre と Rijke の研究では言語リンクで結ばれたページを利用することにより，パラレルコーパス（parallel corpus）を構築する研究を進めている [Adafre 06].

4. チャレンジの現状

筆者らは，本チャレンジの目標である実用的な基盤リソースの構築を実現するために，特別 Web サイト「Wikipedia-lab.org」を立ち上げ，研究成果の公開や情報共有に努めてきた．以降，チャレンジの現状と社会的意義・今後の展開について詳述する．

4·1 基盤リソースの提供

Wikipedia マイニングに関する研究を俯瞰すると，アンカーテキストやバックワードリンク数，インフォボックスなど，共通に利用されている情報があることに気づく．これらの情報は，Wikipedia の標準のダンプデータでは提供されていないため，現状では各研究者がそれぞれ独自に解析・用意している．しかし，Wikipedia 構文をパースするコードを各研究者が個別に用意することはコスト・

精度・標準化の点で効率的ではない．そのため，本チャレンジでは，このように各研究者が共通に使う情報を抽出するスクリプト「Wikipedia RR（Research Resources）」を基盤リソースとして提供している．これは，Wikipediaのダンプデータからアンカーテキストやバックワードリンク数，インフォボックスなどの共通に利用されるデータを抽出し，データベースへ格納するライブラリである．特徴としては，MediaWiki（Wikipedia の CMS）のパーサを内部的に利用してパースしているため，独自コードを書くより精度が高く，例外的な記述方法に強いという特徴を持っている．

本チャレンジでは，個別の研究や開発も進めるが，Wikipedia RR のように各研究者が共通に必要とするデータやリソース，ライブラリの提供も行う．これは，これらのリソースが Wikipedia マイニング研究全体の底上げになり，最終的にはチャレンジの目標達成に貢献するという考えに基づくものである．以降解説する筆者らの研究成果の多くは，Wikipedia RR のデータを基盤リソースとして利用している．

4·2 連想関係抽出

Wikipedia は Wiki をベースにしており，記事の中に他の概念（を意味する単語）が出現するとその概念に対してリンクが張られるため，全体としてみると，概念をノード，ハイパーリンクをリンクとした一種のネットワークと見做すことができる．通常の Web サイトと異なり，ノード（ページ）は概念を表し，リンクは意味的な関係を表す上に，Wikipedia 内部で概念同士が密なリンク構造を形成している（内部リンクが多い）ため，リンク構造を解析することで概念間の関係性を抽出することが可能である．

この特徴を生かし，リンクの構造を解析して関連度計算を行うのがページ間リンクの解析手法である．この分類において，主な手法としては，ネットワーク構造における二つの概念のリンク数やホップ数に基づく手法，記事内における概念の共起性に基づく手法などがある．

筆者らの研究では，pfibf [Nakayama 07] という手法を提案し，大規模な連想シソーラスを構築している．この手法では，概念をノードとしたネットワーク構造において，概念間のパスの数が多いほど，またそれらのパスが短いほど，それらの概念が強く関連していると見做す．それに加えて，ある概念が他の概念からリンクを多く張られている（バックワードリンク数が多い）ほど，その概念が一般的な概念であるとして，関連度を弱くする．さらに，[Nakayama 07] では pfibf に加えて，フォワードリンクによる解析とバックワードリンクによる解析の比重を可変にする手法も提案している．これは，一般的な概念と専門的な概念では，記事の信頼度やフォワードリンク・バックワードリンクの重要性が異なるためである．この手法では，従来手法である TF-IDF と比べて計算量は増加するが，精度は向上している．

また，共起に基づく研究 [Ito 08] では，記事内に出現するリンクの共起性から概念の関連度を計算している．pfibf はリンク構造を繰り返し探索するという解析方法であったが，この手法では各記事を一通り解析すれば十分であるため，pfibf と同程度の精度を実現しつつも計算量が大幅に改善されている．

これらの連想関係抽出の結果は，既に Web アプリケーションとして公開しており，自由に連想関係抽出の結果を確認できる [*1]．また，この連想関係抽出の結果を他の研究者が独自のアプリケーションで利用可能なように，XML Web サービスとしても提供している（図 1）[*2]．現状，C#などの.NET 環境および Java からの利用が可能である．さらに，概念同士の連想関係ネットワークを可視化する「Wikipedia Thesaurus Visualizer」も提供している．図 2 に Wikipedia-lab が提供する基盤リソース・API とアプリケーションの関係を示す．

4·3 関係抽出

前述の通り，従来の Wikipedia マイニング研究では，関係抽出にインフォボックスやカテゴリ，テキスト情報を使った研究が多い．ここで，筆者らはリンク構造解析とテキスト解析の手法を組みあわせた研究を進めている [Nakayama 08]．本研究では，事前にリンク構造解析により重要な単語を抽出しておくことで，テキスト解析に基づく関係抽出の精度とパフォーマンスを向上させることができることを示した．これは，Wikipedia に存在する多量のテキストを全て解析するのは効率的ではなく，ある概念（ページ）にとって重要な概念を含む文章を集中的に解析することで１）不要なパージング処理の回避，２）照応解析精度の向上を果たすものである．たとえば，「Microsoft Windows」というページに存在する二つの文章「Microsoft Windows is the name of several families of software operating systems by Microsoft. 」と２）「The dual routes have generally led to home versions having greater multimedia support and less functionality in networking and security, …」について考える（下線はリンク）．前者は「Microsoft Windows」に対して重要な単語である「Operating System」や「Microsoft」が含まれており，内容もそのページの主題と別概念との関係を示した内容である．しかし，後者の文章では，「Microsoft Windows」の機能の一部分の機能と別概念との関係を示したものであり，そのページの主題とは直接関係しない．このように，関係度の低い概念が出現する文章の内容は，そのページの主題とは直接関係がない場合が多く，このような文章の解析をスキップすることでパフォーマンスと照応解析の精度を向上できることを実験により明らかにした．本手法によって抽出した概念間の関係を表 1 に

∗1 http://wikipedia-lab.org:8080/WikipediaThesaurusV2/
∗2 http://wikipedia-lab.org/ja/index.php/Wikipedia_API

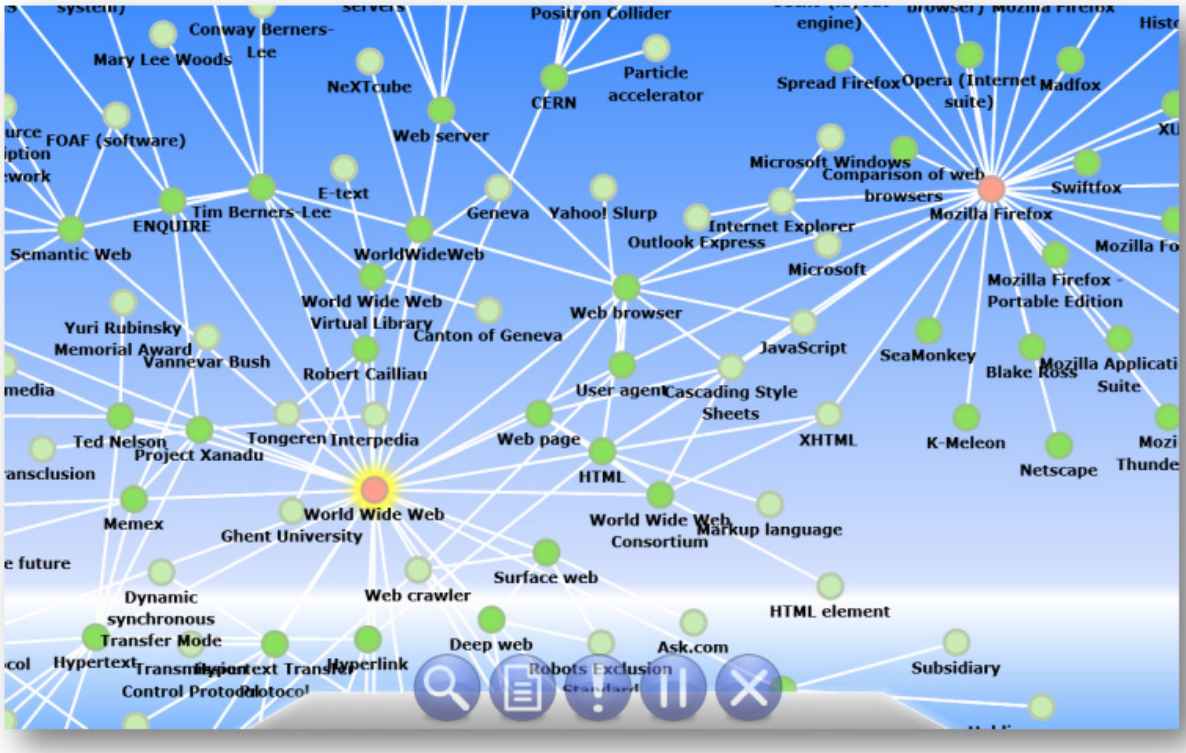

図 1 概念「WWW」に関する連想キーワードの可視化

示す．

4·4 対訳関係抽出

先述のとおり，Wikipedia からの対訳関係抽出では，単に言語リンクを対訳関係として利用していることが多いが，筆者らは言語リンクで網羅されている対訳関係の数に限りがあることを指摘し，言語リンク以外の半構造化情報を利用することで対訳関係が拡張できることを示した [Erdmann 08]．具体的には，アンカーテキストとリダイレクトリンクを解析することで同義語が抽出できることと，バックワードリンク数が情報の信頼性を数値化するのに有効であることを利用した重みづけ関数を定義し，スコア付きの対訳関係を抽出するアルゴリズムを提案している．

また，[Erdmann 09a, Erdmann 09b] では，対訳関係抽出の精度や網羅性を向上させるためにはどのような素性が有効かを列挙し，SVM によって学習することでより実用性の高い対訳関係辞書を作成している．

4·5 今後の展開

近未来チャレンジ「Wikipedia マイニング」では，Wikipedia を解析することで，幅広い分野で利用可能な基盤リソースを構築していくことを目的としている．基盤リソースには，「連想シソーラス」「関係辞書」「対訳辞書」が含まれる．さらに，これらの基盤リソースへアクセスするための API（Web サービス）や解析用のリソースなどを提供していくことで，他研究者や開発者が効率良く研究開発ができる環境を築く．これは，本分野で質の高い研究が行われれば，さらにそれらの知見を基に本チャレンジの質を向上できるからである．

本チャレンジによって質の高い基盤リソースが構築できた際の社会的意義は大きく，情報検索・機械翻訳・テキスト分類などの幅広いアプリケーションを高度化できると期待される．たとえば，情報検索においては連想関係を利用して，関係する文書を提示することや，概念同士の関係を利用して必要な情報と不要な情報を識別することなどが可能となる．さらに，対訳辞書を使うことで，多言語で記述された情報同士をリンクすることが可能となる．Wikipedia マイニングによる対訳抽出は，専門分野の概念や新しい概念に対する網羅性が高いため，Web のように多様な情報が存在する環境では有効だと考えられる．

これらのビジョンを達成するためには，いくつかの技術的な課題が残されているが，本チャレンジで特に注力

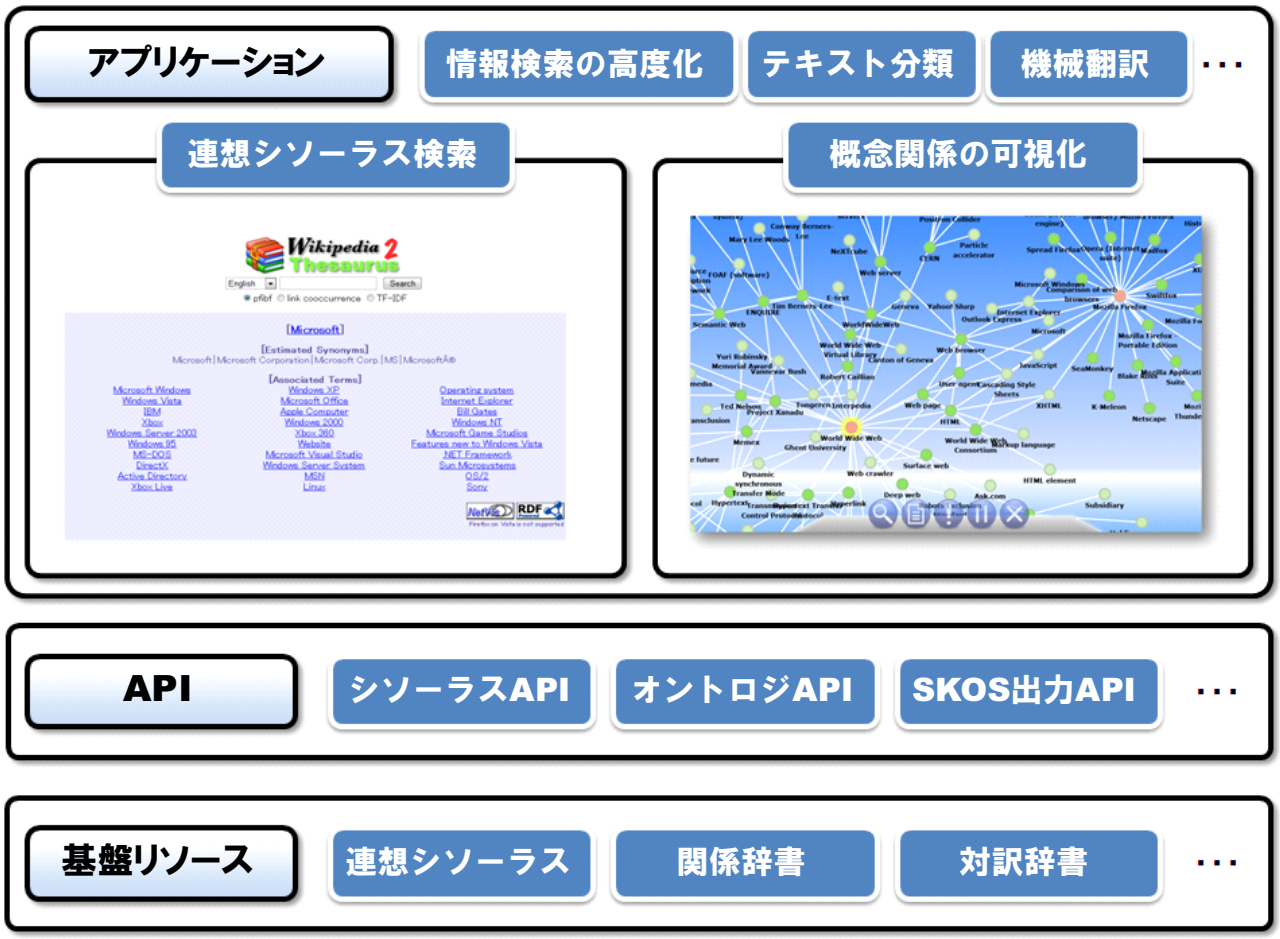

図 2 Wikipedia-lab が提供する基盤リソース・API とアプリケーションの関係

するのは，1）スケーラビリティの高い解析手法の確立と 2）リンク構造と NLP の融合である．1）のスケーラビリティに関しては，Wikipedia のような大規模データの解析においては小さな問題ではなく，アルゴリズム設計に大きな影響を与える．この点に関しては，pfibf で利用している効率の良いリンク解析手法や，重要文解析による計算量の削減などが有効であるといえる． 2）のリンク構造解析と NLP の融合も重要な課題の一つである．Wikipedia はフリーテキスト以外にもカテゴリやページ間リンク，対訳リンクなど，情報同士をリンクする仕組みが豊富に用意されている．これらの情報を補助的に使うことにより，コンテキストの認識や関係度の抽出を通じて，NLP の精度向上が期待できる．

また，Wikipedia から抽出した基盤リソースの有効性を証明するため，本チャレンジでは 5 年以内に基盤リソースを使った実アプリケーションを構築する予定である．目標として「汎用的な基盤リソースの構築」を掲げているので，アプリケーションとしては多様な方向性が考えられるが，そのなかでも特に情報検索の高度化に注力していきたい．具体的には，連想関係を利用した関連文書検索，関係辞書を利用した意味ベースの情報検索，対訳辞書を利用した異言語横断文書検索などを実現することを目指す．

5. ま と め

本稿では，近未来チャレンジ「Wikipedia マイニング」の概況および今後の展開について解説した．Wikipedia は，意味情報を中心とした WWW を実現するための基盤リソースとして注目されている．本チャレンジでは，今後「連想関係抽出」「関係抽出」「対訳抽出」の 3 点に注力して研究を進めていく．

連想関係抽出に関してはカテゴリツリーの解析，ページ間リンクの解析，コンテンツ類似性などの種々の指標があるが，今後はこれらの指標を融合した手法や，他リソースとの融合が重要になると考えられる．これらの解析結果を「Wikipedia Thesaurus」に適用し，公開していく予定である．関係抽出やオントロジの分野では，ブーストラッピング手法による関係の抽出が主流になりつつあるが，リンク構造解析手法を取り入れるなど，NLP 以外の手法やリソースを利用することでパフォーマンスの向上が期待できる．言語間リンクの研究に関しては，単に言語間リンクを抽出して何らかのタスクによって評価した，といったレベルの研究が多く，まだ多くの研究的課題が残っているといえる．本チャレンジでは，既に「Wikipedia Translation Dictionary」を公開しているが，精度・網羅性を向上させる取り組みを続けており，順次公開システ

表 1 関係抽出の結果

LSP 法（リードセンテンス解析法）の結果

Subject	Predicate	Object
Apple	is-a	Fruit
Bird	is-a	Homeothermic
Cat	is-a	Mammal
Computer	is-a	Machine
Isola d'Asti	is-a	Comune
Jimmy Snuka	is-a	Pro. wrestler
Karwasra	is-a	Gotra
Mineral County	is-a	County
Sharon Stone	is-a	Model
Sharon Stone	is-a	Film producer

ISP 法（重要文解析法）の結果

Subject	Predicate	Object
Odonata	is an order of	Insect
Clarence Thomas	was born in	Pin Point, GA
Dayton, Ohio	is situated	Miami Valley
Germany	is bordered on	Belgium
Germany	is bordered on	Netherlands
Mahatma Gandhi	founded	N. Indian Congress
Mahatma Gandhi	established	Ashram
Rice	has	Leaf
Rice	is cooked by	Boiling
Rice	is cooked by	Steaming

ムに適用していく予定である．

謝　辞

本研究の一部は，マイクロソフト産学連携研究機構 CORE 連携研究プロジェクトの助成と，科学研究費補助金（基盤研究 (B) 21300032, 基盤研究 (C) 20500093) によるものである．ここに記して謝意を表す．

◇ 参 考 文 献 ◇

[Adafre 06] Adafre, S. F. and Rijke, de M.: Finding Similar Sentences across Multiple Languages in Wikipedia, in *Proc. of the EACL Workshop on NEW TEXT Wikis and blogs and other dynamic text sources*, pp. 62–69 (2006)

[Auer 07] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. G.: DBpedia: A Nucleus for a Web of Open Data, in *Proc. of International Semantic Web Conference, Asian Semantic Web Conference (ISWC/ASWC)*, pp. 722–735 (2007)

[Bouma 06] Bouma, G., Fahmi, I., Mur, J., Noord, van G., Plas, van der L., and Tiedemann, J.: The University of Groningen at QA@CLEF 2006 Using Syntactic Knowledge for QA, in *Working Notes for the Cross Language Evaluation Forum Workshop (CLEF)* (2006)

[Culotta 06] Culotta, A., McCallum, A., and Betz, J.: Integrating Probabilistic Extraction Models and Data Mining to Discover Relations and Patterns in Text, in *Proc. of Human Language Technology Conference - North American Chapter of the Association for Computational Linguistics (HLT-NAACL)* (2006)

[Erdmann 08] Erdmann, M., Nakayama, K., Hara, T., and Nishio, S.: An Approach for Extracting Bilingual Terminology from Wikipedia, in *Proc. of International Conference on Database Systems for Advanced Applications (DASFAA)* (2008)

[Erdmann 09a] Erdmann, M., Nakayama, K., Hara, T., and Nishio, S.: Improving the Extraction of Bilingual Terminology from Wikipedia, *ACM Transactions on Multimedia Computing, Communications and Applications* (2009)

[Erdmann 09b] Erdmann, M., Nakayama, K., Hara, T., and Nishio, S.: Using an SVM Classifier to Improve the Extraction of Bilingual Terminology from Wikipedia, in *Proc. of WikiAI conjunction with International Joint Conference on Artificial Intelligence* (2009)

[Ferrández 07] Ferrández, S., Toral, A., Ferrández, Óscar , Ferrández, A., and noz, R. M.: Applying Wikipedia's Multilingual Knowledge to Cross-Lingual Question Answering, in *Proc. of International Conference on Applications of Natural Language Processing and Information Systems (NLDB)*, pp. 352 – 363 (2007)

[Gabrilovich 07] Gabrilovich, E. and Markovitch, S.: Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis, in *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1606–1611 (2007)

[Haghighi 08] Haghighi, A., Liang, P., Berg-Kirkpatrick, T., and Klein, D.: Learning Bilingual Lexicons from Monolingual Corpora, in *Proc. of Association for Computational Linguistics (ACL)* (2008)

[Ito 08] Ito, M., Nakayama, K., Hara, T., and Nishio, S.: Association Thesaurus Construction Methods based on Link Co-occurrence Analysis For Wikipedia, in *Proc. of ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 817–826 (2008)

[Leuf 01] Leuf, B. and Cunningham, W.: *The Wiki Way: Collaboration and Sharing on the Internet*, Addison-Wesley (2001)

[Li 03] Li, X. and Liu, B.: Learning to Classify Texts Using Positive and Unlabeled Data, in *Proc. of International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 587–594 (2003)

[Nakayama 07] Nakayama, K., Hara, T., and Nishio, S.: Wikipedia Mining for An Association Web Thesaurus Construction, in *Proc. of International Conference on Web Information Systems Engineering (WISE)*, pp. 322–334 (2007)

[Nakayama 08] Nakayama, K., Hara, T., and Nishio, S.: Wikipedia Link Structure and Text Mining for Semantic Relation Extraction, in *Proc. of Semantic Search Workshop (SemSearch)*, pp. 59–73 (2008)

[Nguyen 07] Nguyen, D. P. T., Matsuo, Y., and Ishizuka, M.: Relation Extraction from Wikipedia Using Subtree Mining, in *Proc. of Conference on Artificial Intelligence (AAAI)*, pp. 1414–1420 (2007)

[Pantel 06] Pantel, P. and Pennacchiotti, M.: Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations, in *Proc. of International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (COLING/ACL)* (2006)

[Ponzetto 07] Ponzetto, S. and Strube, M.: Deriving a Large Scale Taxonomy from Wikipedia, in *Proc. of Conference on Artificial Intelligence (AAAI)*, pp. 1440–1447 (2007)

[Schönhofen 07] Schönhofen, P., Benczúr, A., Bíró, I., and Csalogány, K.: Performing Cross-Language Retrieval with Wikipedia, in *Working Notes for the Cross Language Evaluation Forum Workshop (CLEF)* (2007)

[Strube 06] Strube, M. and Ponzetto, S.: WikiRelate! Computing Semantic Relatedness Using Wikipedia, in *Proc. of Conference on Artificial Intelligence (AAAI)*, pp. 1419–1424 (2006)

[Suchanek 07] Suchanek, F. M., Kasneci, G., and Weikum, G.: YAGO: A Core of Semantic Knowledge, in *Proc. of International Conference on World Wide Web (WWW)*, pp. 697–706 (2007)

[Suchanek 08] Suchanek, F. M., Kasneci, G., and Weikum, G.: YAGO: A Large Ontology from Wikipedia and WordNet, *Journal of Web Semantics*, Vol. 6, No. 3, pp. 203–217 (2008)

[Wang 07] Wang, G., Yu, Y., and Zhu, H.: PORE: Positive-Only Relation Extraction from Wikipedia Text, in *Proc. of International Semantic Web Conference, Asian Semantic Web Conference (ISWC/ASWC)*, pp. 580–594 (2007)

[Wikimedia Foundation 08] Wikimedia Foundation, : CategoryTree., *http://en.wikipedia.org/wiki/Special:CategoryTree* (2008)
[中山 07] 中山 浩太郎, 原 隆浩, 西尾 章治郎：人工知能研究の新しいフロンティア：Wikipedia(アーティクル), 人工知能学会誌, Vol. 22, No. 5, pp. 693–701 (2007)

〔担当委員：阿部 明典〕

2009 年 2 月 7 日 受理

著 者 紹 介

中山 浩太郎(正会員)

2001 年関西大学総合情報学部卒業．2003 年同大学院総合情報学研究科修士課程修了．この間 (株) 関西総合情報研究所代表取締役社長，同志社女子大学非常勤講師に就任．2004 年関西大学大学院を中退後，2007 年大阪大学大学院情報科学研究科にて博士号を取得し，同年 4 月から大阪大学大学院情報科学研究科特任研究員，2008 年 4 月から東京大学 知の構造化センター特任助教に就任し，現在に至る．人工知能および WWW からの知識獲得に関する研究に興味を持つ．IEEE，ACM，情報処理学会，電子情報通信学会の各会員．

伊藤 雅弘(学生会員)

2007 年立命館大学理工学部情報学科卒業．2008 年大阪大学大学院情報科学研究科博士前期課程修了．同年 4 月から大阪大学大学院情報科学研究科博士後期課程に進学，現在に至る．人工知能，WWW からの知識獲得および情報検索に関する研究に興味を持つ．情報処理学会，日本データベース学会の各学生会員．

Erdmann, Maike

2006 年ドイツ・CvO 大学卒業．2008 年大阪大学大学院情報科学研究科博士前期課程修了．同年 4 月から大阪大学大学院情報科学研究科博士後期課程に進学，現在に至る．自然言語処理，WWW からの知識獲得に関する研究に興味を持つ．

白川 真澄

2008 年大阪大学工学部電子情報エネルギー工学科卒業．同年 4 月から大阪大学大学院情報科学研究科博士前期課程に進学，現在に至る．人工知能，WWW からの知識獲得および情報検索に関する研究に興味を持つ．情報処理学会，日本データベース学会の各学生会員．

道下 智之

2008 年神戸大学工学部情報知能工学科卒業．同年 4 月から大阪大学大学院情報科学研究科博士前期課程に進学，現在に至る．自然言語処理，WWW からの知識獲得および情報検索に関する研究に興味を持つ．

原 隆浩

1995 年大阪大学工学部情報システム工学科卒業．1997 年同大学院工学研究科博士前期課程修了．同年，同大学院工学研究科情報システム工学専攻助手，2002 年同大学院情報科学研究科マルチメディア工学専攻助手，2004 年より同大学院情報科学研究科マルチメディア工学専攻准教授となり，現在に至る．工学博士．2000 年電気通信普及財団テレコムシステム技術賞受賞．2003 年情報処理学会研究開発奨励賞受賞．2008 年，2009 年情報処理学会論文賞．IEEE，ACM，電子情報通信学会，日本データベース学会の各会員．

西尾 章治郎(正会員)

1975 年京都大学工学部数理工学科卒業．1980 年同大学院工学研究科博士後期課程修了．工学博士．京都大学工学部助手，大阪大学基礎工学部および情報処理教育センター助教授，大阪大学大学院工学研究科情報システム工学専攻教授を経て，2002 年より大阪大学大学院情報科学研究科マルチメディア工学専攻教授となり，現在に至る．2000 年より大阪大学サイバーメディアセンター長，2003 年より大阪大学大学院情報科学研究科長，その後 2007 年より大阪大学理事・副学長に就任．本会評議員を歴任．本会論文賞を受賞．