社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－DBS－145 (27)
2008－ＦＩ－ 91 (27)
2008／6／20

Wikipediaマイニングと Webオントロジ構築の新しい方向性

中山浩太郎† 原隆浩‡ 西尾章治郎‡
東京大学知の構造化センター† 大阪大学大学院情報科学研究科‡
nakayama@cks.u-tokyo.ac.jp hara@ist.osaka-u.ac.jp nishio@ist.osaka-u.ac.jp

Wikipedia Mining and a New Paradigm on Web Ontology Construction

Kotaro NAKAYAMA† Takahiro HARA‡ Shojiro NISHIO‡
The Center for Knowledge Structuring, the University of Tokyo†
Dept. of Multimedia Eng., Graduate School of Information Science and Technology, Osaka University‡

1. はじめに

Wikipediaは，Wikiをベースにした大規模Web百科事典であり，誰でもWebブラウザを通じて記事内容を変更できることから，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200万（2007年12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興味深い特徴をいくつか持つ．密なリンク構造や，URLにより語彙の意味が一意に特定されている点，質の高いリンクテキスト，精査されたカテゴリ構造や情報ボックス（Infobox）などもWikipediaの知識抽出のコーパスとして見たときの特徴である．筆者らはWikipediaの特徴を利用して，概念同士の関係度を数値化した連想シソーラスの構築に関する研究を進め，その実装である「Wikipediaシソーラス」(http://wikipedia-lab.org参照)を公開した．連想シソーラスは，概念間の関係度を数値化した辞書であり，情報検索や自然言語処理，情報フィルタリングなど幅広い分野で必要とされている．しかし，連想シソーラス辞書は，単に概念間の関係度を数値化しているだけであり，より明確な意味関係を抽出することで大規模Webオントロジを構築することが本研究の目標である．Wikipediaからの Webオントロジ構築はセマンティックWeb研究や概念辞書構築の新しいパラダイムとなる可能性が高く，国内外で注目を集めている研究分野である．

2. 技術的課題

前述のとおり，Wikipediaは知識抽出のコーパスとして有用であるが，①情報の信頼性，②スケーラビリティ，③情報統合などの技術的課題も残されている．特に，情報の信頼性に関する問題は，Wikipediaのように多数のユーザがコンテンツを編集するソーシャルメディアでは重要な問題であり，質の高い情報だけを抽出できるかがポイントになる．また，Webオントロジ構築は，単体だけでは意味がなく，アプリケーションやタスクを考慮して手法を開発・最適化する必要がある．たとえば，情報検索の分野では比較的精度よりも網羅性が必要とされる場合が多いが，機械翻訳のための要素技術として利用する場合は，精度が極めて重要になる．

3. 研究の方向性

本研究ではWikipediaマイニングによって，概念間の明確な意味関係（is-a関係など）を定義したWebオントロジを自動構築する手法の構築を目指す．また，その際に上述の問題点を解決する基盤技術を開発していくことが研究のポイントとなる．

①情報の信頼性：Wikipediaの記事内容の信頼性を数値化するためのパラメータとして利用できる指標はいくつかある．バックワードリンク数や記事の更新状況，生存期間などがその最たる例である．他にどのような指標がWikipediaの記事の信頼性に影響を与えるのかを調査・利用することで質の高い情報抽出を目指す．

②スケーラビリティ：Wikipediaのテキストを解析するには，構文解析が必要となるが，構文解析は多量の計算機リソースを消費するため，日に日に増加するWikipedia記事をすべて解析するためには計算量を抑える仕組みが必要となる．そのため，1）インクリメンタルな解析手法，2）重要文だけの解析によるパフォーマンス向上，3）非構文解析手法の検討などの研究を行う必要がある．

③情報統合：Wikipediaは専門的な概念や新しい単語に対する網羅性が高い．その一方で，概念辞書であるWordNetやOpenCYCなどは一般的な概念がよく網羅されているため，融合することで両者の利点を活かした概念辞書が構築できると考えられる．また，Webマイニング手法との融合手法により，さらに網羅性を高めるつもりである．

④アプリケーションへの応用：本研究では，開発した大規模シソーラスおよび大規模オントロジを公開している．これらのリソースをほかの研究者に公開することで，アプリケーションへの応用の可能性を促進し，より有用なWebオントロジの構築を目指す．

参考文献

1) 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，人工知能学会誌，Vol. 22, No. 5 (2007年9月).