【15】

氏　　名	中山　浩太郎（なかやま　こうたろう）
博士の専攻分野の名称	博　士（情報科学）
学位記番号	第　２１３２１　号
学位授与年月日	平成 19 年 3 月 23 日
学位授与の要件	学位規則第４条第１項該当 情報科学研究科マルチメディア工学専攻
学位論文名	**リンク構造解析に基づく大規模 Web コンテンツからの知識獲得に関する研究**
論文審査委員	（主査） 教　授　西尾章治郎 （副査） 教　授　藤原　　融　　教　授　岸野　文郎　　教　授　薦田　憲久 教　授　下條　真司　　助教授　原　　隆浩

論　文　内　容　の　要　旨

本論文は、筆者が 2004 年から現在までに、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在学中に行った研究をまとめたものである。

1990 年代初頭に初期の WWW が登場してから十数年経過した現在、次世代 Web の実現の動きが活発化してきている。WWW の発明者である Tim Berners-Lee が率いるセマンティック Web の活動などに始まり、大規模 Web 事典（Wikipedia）、フォークソノミー（Flickr、Del. icio. us）、ソーシャルネットワーク（Friendster、MySpace）など、WWW の構造に大きなパラダイムシフトが発生している。

次世代の WWW においては、セマンティック Web のように情報の意味そのものを取り扱う情報共有基盤が必要とされている。これを実現するためには、機械可読な意味データが重要な役割を果たし、このようなデータとして Web オントロジが注目されている。オントロジは、概念と概念同士の関係を明確に定義した一種の辞書である。しかし、このような概念辞書は、人の手によって構築するためには多大な労力と時間を要する上に、規模や更新頻度といった問題が発生する。そのため、これらの機械可読の概念辞書を構築し、WWW を通じて共有するための技術が必要とされている。

これまでに、通常のテキストや Web コーパスから有用な知識を自動的に抽出し、機械可読な辞書を構築する研究は数多く行われてきたが、それらのほとんど全てが自然言語処理を利用するものであった。自然言語処理には、同義語・多義語の問題や形態素への分割など、語と語の関係を調べる前の段階で精度低下の要因がいくつか存在し、技術的な課題として未だ解決されていない。

そこで、本研究では Web コンテンツがトピック局所性などの重要な属性を持ち、リンクテキスト、バックワードリンク、コンテンツのトポロジなどを解析することで有用な情報を抽出できることに着目し、大規模 Web コンテンツを解析することで有用な情報を抽出する手法について検討した。

本論文は、５章から構成され、その内容は次のとおりである。まず、第１章において序論として研究の背景を述べた。次に、第２章では、大規模 Web 事典のリンク構造を解析し、自然言語処理による精度低下を防ぐことで、精度の高いシソーラス辞書（連想概念辞書）を構築する方法を提案した。本手法は、有向グラフ内のすべての要素ペアに

対して関連性を求めるアルゴリズム lfibf（Link Frequency Inversed Backward link Frequency）と、Web 事典のリンク構造の特性を考慮し、フォワードリンクとバックワードリンクの重み付けを変更するアルゴリズムをはじめとした３つの応用手法を提案し、シソーラス辞書の精度向上を目指すものである。また、数百万の要素数を持つ Web 事典に最適化した大規模隣接推移確率行列の圧縮方式と解析アルゴリズムを開発した。その結果、膨大な量のページを持つ Web コンテンツに対しても更新に応じて辞書更新が可能となった。さらに、提案方式の性能評価のために行った実験の結果を示し、その有効性について検証した。

第３章では、上記の手法によって構築された精度の高い連想シソーラスを Web オントロジへ昇華させるために「パーソナルオントロジ」の提案とその構築方法を提案した。次世代 WWW においては、Web オントロジが情報検索のために極めて重要な役割を果たすが、前述のとおり自動構築時の精度低下が問題であった。そのため、本研究では、大規模 Web シソーラスと概念コーパス（WordNet、Mindpixel など）をマッチングする手法を提案し、シソーラスと概念コーパスを統合することで人間のコモンセンスに近い強度付きオントロジの自動構築を目指した。さらに、提案方式の性能評価のために行った実験の結果を示し、その有効性について検証した。

第４章では、情報の信頼性についての考察を行った。具体的には、ソーシャルネットワーク上での友人関係を解析することで、信頼関係を数値化するアルゴリズムを提案し、コンテンツの信頼性を実現した。さらに、アプリケーションとして、地理情報共有システム GeoNote.net を実際に構築・運営し、ソーシャルネットワーク構築および地理情報検索に関する実験を行うことで、実際の環境に即した状況で信頼性アルゴリズムを評価した。

最後に第５章では、本論文の成果を要約したのち、今後の検討課題について述べ、本論文のまとめとした。

論文審査の結果の要旨

本論文は、近年活発化してきた次世代 Web 技術に着目し、大規模な Web コンテンツから有用な知識を抽出するための手法についてまとめたものである。その主要な成果を要約すると次の通りである。

⑴大規模 Web 辞典からシソーラス辞書を自動的に構築する手法として lfibf と３つの応用手法「単純法」「対数近似法」「FB 法」を提案している。また、スケーラビリティを確保するために、大規模行列積を計算するためのデータ構造として二重二分木を提案している。評価実験では、精度、スケーラビリティともに従来のシソーラス辞書構築手法の問題を解決できていることを証明している。また、自然言語とのマッチングを行うことで、情報検索などへの応用も可能となっている。

⑵上記手法によって構築されたシソーラスと語彙体系のコーパスをマッピングすることで、パーソナライズされた概念体系の構築方法を提案している。また、そのパーソナルオントロジの表現方法およびその構築方法も提案している。

⑶ソーシャルネットワークにおける友人関係を解析し、信頼性を数値化するアルゴリズムを提案している。このアルゴリズムの実問題への応用として地理情報共有システムへ適用し、評価実験によりその有用性を証明している。

以上のように、本論文は次世代 Web 技術の特性に着目し、大規模な Web コンテンツからの知識抽出に関する先駆的研究として、情報科学に寄与するところが大きい。よって、本論文は博士（情報科学）の学位論文として価値のあるものと認める。