審 査 の 結 果 の 要 旨

氏　　名　　大澤　昇平

この学位請求論文「ウェブの情報統合に基づく人物属性の推定の方法論に関する研究」では、ウェブの情報に基づき、複数の情報源から情報を統合することで、人物の能力や人気等の人物属性を推定する方法について論じている。

近年、ウェブの進展やソーシャルメディアの発達にともない、さまざまなデータを取得し、活用することが可能になっている。例えば、Github というサービスでは、ウェブのエンジニア（あるいはプログラマ）の活動がオープンに記録されており、これを用いることで、どのエンジニアにどのようなスキルがあるのかを推定することができる。DBpedia というデータでは、概念間の関係性が、コンピュータが容易に利用可能な形で提供されている。本論文では、こうしたデータを複数組み合わせることで、機械学習の素性をうまく設計し、精度よく人物属性を推定しようというものである。

本論文は、大きく３つの個別研究から構成される。ひとつは、人物属性を推定するための情報取得に関する研究である。近年のウェブサービスは、API（Application Program Interface）と呼ばれる仕組みを備え、外部からサードパーティーがデータを活用することを促進している。ただし、サーバの負荷等を考えて、一定の制限があることが普通である。こうした制限のなかで、目的に応じたデータを取得することは、現実的な必要性が高いと同時に、学術的にも新しい問題を提示している。この問題の本質は、できるだけサンプリングの数を抑えながら、目的に応じて対象をできるだけカバーすることであり、新たなサンプリング手法が求められる。このため、本論文では、辞書ベースサンプリングという手法を提案し、複数の辞書のなかで、どの辞書を用いてデータを取得するのが目的に最も合致するのかを測定し、その辞書を使って情報の取得を行う。辞書のなかの単語を使って、API に検索のクエリーを出し、データをサンプリングするという手続きである。結果として、提案手法により、Facebook ページを効率的にサンプリングすることが可能となった。

個別研究の２つ目は、エンジニアの能力推定に関するものであり、前述の Github のデータ、および oDesk というクラウドソーシングのデータを組み合わせて用いている。oDesk におけるエンジニアに対する評価を、Github のデータに由来する h-index 等の指標により的確に取得することができることを示した。

また、Github のデータと oDesk のデータの両方を用いることで、エンジニアの能力の分析も行っている。Github のデータの分析はこれまでほとんど行われておらず、ニーズの高いエンジニアの能力の分析を実データで行っているという点からも、意義のある研究であると言える。

個別研究の３つ目は、Facebook の Like 数を予測しようというもので、人物の人気を予測することを行っている。そのために、Facebook ページのうち人物に関するものを集め、それをエンティティマッチングと呼ばれる技術により、DBpedia のデータと紐付けている。人物の人気を予測するために、DBpedia で意味的に関連づけられたエンティティの情報を用いる。意味的に関連するエンティティの Like 数の合計や平均等を機械学習の素性として用いることで、従来手法よりも精度よく Like 数が予測できることを示した。特に、エンティティマッチング、および意味的な情報からの機械学習の素性生成という点について、新規性のある研究である。

以上の３つの研究を通じて、本研究では、ウェブ上の情報を統合することで、人物の属性を推定することが可能であることを示した。本研究で示した手法は、本質的には人物に限らず、さまざまなエンティティに適用可能である。その点で有用性は高い。また、これまでにウェブ上の情報を統合する研究は行われており、例えばマッシュアップと呼ばれる技術では、複数の情報をウェブサービスとして統合するものであった。しかし、本研究では、ウェブ上の複数の情報を、機械学習の素性生成に焦点を当てて統合することを意図しており、その点で新規性が高い。ビッグデータの活用が進んでいる時代背景にもあった提案であり、ウェブ上のデータの新たな活用方法を提案するものでもある。その際、API の制限などの現実的な制約を踏まえる必要があり、本研究では、個別研究の１つとして、API の制限のみを対象として課題解決を図っているが、そういった現実的な制約を解消していく必要があることを示したことも、本研究の重要な示唆のひとつである。

本論文では、人物属性の推定という具体的な課題を通じて、機械学習の素性生成のための情報統合の方法論を提案しており、ウェブ工学の分野において重要な貢献である。また社会学や経済学からみても、新しいシーズ技術としての可能性を秘めており、その可能性を示したという点でも、本論文の意義は大きい。本論文は、新しい技術の適用可能性を示すと同時に、有用な知見を提供しており、博士（工学）の学位論文としてふさわしい。

よって本論文は博士（工学）の学位請求論文として合格と認められる。