博士論文（要約）

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

指導教員　松尾 豊 准教授

東京大学大学院工学系研究科

技術経営戦略学専攻

大澤 昇平

平成 26 年 12 月1 日提出

内容梗概

本研究では，ウェブの情報に基づき，機械学習を用いて人物属性の推定を行う．人物属性とは性別や能力などの人物の特徴のことである．人物属性の推定は，採用活動やマーケティングなど対象とする人物のことを知る必要がある活動において有用である．

機械学習の枠組みにより人物属性の推定を行なう研究に関しては，Twitter から性別を推定する研究などが挙げられる．しかし，ウェブ上の情報は欠損が多く，単一の情報源からでは知ることのできない情報も多い．一方，機械学習の枠組みにより人物属性の推定を行う研究として，分散する複数の情報源を統合するという観点での研究については行われていない．こうした研究には，エンティティ・リンキングなどの情報統合の知見を活かすことができると考えられる．複数の情報源を用いることで属性推定の精度向上を行なうことが期待できる．

こうした研究が行われていない理由は，最近まで，ウェブ上に公開されている人物情報は企業が特定の目的のみに利用することを規定しており，そこから分析者が情報統合を目的としてデータを取り出すことは困難だったからである．しかし，近年，一部の企業やパブリックセクターが人物情報をウェブ上の二次利用を促進する仕組みを通してデータを提供する事例が増えてきており，これらから得られる情報を活用できることが期待できる．このような仕組みから得られる情報を統合し属性推定を行うために，本研究では，a) 属性

推定に活用できる情報の収集法，b) 収集した情報を統合した素性生成の方法について研究を行う．a) では，ウェブ上の各種サービスが提供している検索 API の活用に焦点を当て，b) では，推定する属性として能力および人気度のふたつを例にとる．

具体的な課題点として，以下のことが挙げられる．

- 属性の推定に活用できる情報を収集するために，各種ウェブサービスの検索 API を用いる必要があるが，なるべく多くの情報を取得するために効率化（少ない回数でなるべく網羅的にサンプリングすること）することが重要である．
- 収集した情報を統合して機械学習の素性を生成し，有効性を示す必要がある．まず，専門家の能力の推定のための素性生成を行う．
- 同様に，属性推定のための有効な情報源である Linked Open Data に含まれる，エンティティ間の意味的関係を統合し，機械学習を適用する．意味的関係を統合するときには，複数のエンティティ（人物や組織等）の対応関係を取る，エンティティ・リンキングが必要となる．SNS における人気度の推定を例にとる．

これらを解決するために，本研究では次のことを行なった．

- 検索 API からのサンプリングの効率化を行なうために，検索 API に入力するクエリを，何らかの辞書を用いて効率化する，辞書ベースサンプリングを提案し，有効性を示した．
- オープンソースコミュニティのひとつである GitHub を対象に，エンジニアの開発物から素性生成を行い，機械学習を適用することで，エンジニアの能力の推定を行うことができることを示した．
- Facebook を対象に，意味的関係として Wikipedia（正確には DBpedia）を組み合

わせ，エンティティ・リンキングを行った上で素性生成を行い，機械学習を適用することで，Facebook における人気を推定することができること，また，意味的関係を統合する効果があることも示した．

検索 API からのサンプリングでは，実際に提案手法を用いて Facebook からのサンプリングを行った結果，目的とするエンティティ集合のうち約 4 分の 1 をサンプリングできたことが明らかになった．これは，当時の基準で言えば，研究者がオープンデータから手に入れられるものの中で，世界で最も多い Facebook ユーザのデータセットであった．GitHub に蓄積された専門家の開発物に基づく能力推定では，クラウドソーシングの情報を統合することで，実績に基づく能力予測が可能になった．意味的関係を用いた人気度の推定では，個人の人気を高めるのに重要な要因として，上から順に仕事，両親，受賞歴であることが明らかになった．

このような要素課題に関する研究を通し，機械学習により人物属性の推定の手法を構築する方法論ができたといえる．