氏　　名	**上田　洋**
学位の種類	博士（工学）
学位記番号	第5482号
学位授与年月日	平成22年3月24日
学位授与の要件	学位規則第4条第1項
学位論文名	**Web上の人物理解のための人物関連情報の抽出**
論文審査委員	主査　教授　辰巳　昭治　　副査　教授　鳥生　隆
	副査　教授　岡　育生　　副査　教授　村上　晴美

論文内容の要旨

本論文では，Web 上の同姓同名人物の識別を容易にすることを目的に，各人物の Web ページクラスタのラベルとして用いる人物関連情報を Web ページから抽出する手法を検討した．人物関連情報とは，職業関連情報，位置情報，履歴書である．同姓同名人物毎に分けられた Web ページクラスタから，職業関連情報，位置情報，履歴書を抽出・作成する手法を提案した．職業関連情報，位置情報，履歴書のそれぞれの手法について，評価実験を行い，その有効性を示した．

第1章では，本論文の背景，目的と人物の識別に有用な人物関連情報について述べた．

第2章では，職業関連情報の抽出手法について述べた．職業関連情報とは，厳密に職業と定義される語だけではなく，幅広く職業と考えられる語や，職業の推定に有用と思われる語であり，具体的には，職業を表す語，所属と役職を表す語，著作と役割を表す語，を抽出する．HTML 構造に着目した名詞の抽出，簡単なヒューリスティックを用いた職業関連候補判定，出現頻度と同義クラスタ作成と Web 検索エンジンを用いたランキングに基づく職業関連情報作成手法を提案した．職業関連情報の抽出手法の特徴は，厳密に定義された職業を出現頻度に応じて付与するのではなく，文脈に応じて人物の識別に有用と思われる職業関連情報を付与するという着眼点にある．特別な職業辞書が不要であることも利点である．

第3章では，位置情報の抽出手法について述べた．位置情報とは，住所，緯度・経度で構成される地理的情報である．ローカル検索エンジンを利用して位置情報を取得する．位置情報の抽出における主要なアイディアは，位置情報候補を抽出するためのランドマークへの着目，人物に最も関連のある位置情報を付与するための語間距離と Web 検索順位の利用である．位置情報の抽出における特徴は，特別な辞書を必要とせず，簡単な手法の組み合わせにより，Web 上の同姓同名人物クラスタ毎に人物に最も関連のある位置情報を付与できる点である．

第4章では，履歴書の作成手法について述べた．履歴書は，一般的な履歴書を参考に，抽出した履歴文（時間と人物に関する出来事両方を含む文）を戸籍，学歴，経歴，受賞歴のカテゴリ毎に分類する．履歴書の作成手法の特徴は，ヒューリスティックを用いた履歴文の抽出とそのフィルタリング，主に HTML のタグの出現パターンを用いた人物と履歴文の関係性判定，機械学習手法を用いた履歴文のカテゴリへの分類，記述された時間を考慮した同義の履歴文のクラスタリングである．

第5章では，情報抽出手法における関連研究と本研究の位置づけについて述べた．

第6章でまとめを行った．

論文審査の結果の要旨

インターネット技術の普及に伴い，情報検索エンジンを用い，Web 上から様々な情報取得が容易となってきた．Web 検索エンジンを利用する主要な目的の一つに，人物に関する情報取得が挙げられる．Web 検索エンジンを用いて人名による検索（人名検索）を行った場合，同姓同名の多数の人物に関する情報が得られる．これを同姓同名問題と呼び，異なった人物として識別することが求められる．また，人物に関する取得情報を簡便な形で明示する必要がある．本研究では，同姓同名人物問題の解消方法と，人物理解のための簡潔な表現形式への変換方法について考察し，以下のような結果を得ている．

まず，同姓同名問題を解消するため，人物に付随している職業関連情報を抽出し，人物を識別するラベルとする方法を用いている．職業関連情報としては，職業を表す語，所属と役職を表す語、著作

と役割を表す語を定義している．Webページ上で記述されている文章から名詞を抽出し，ヒューリスティックを用いて，これらの職業関連情報の候補を抽出し，職業関連情報の候補の集まりから，人名と関連の高い候補を職業関連情報として選択する方法を提案している．提案した方法の特徴は，特別の職業辞書を用いず，Web上のページから人名と職業関連情報との関連度に基づき，職業関連情報を抽出していることである．評価実験を行い，本提案手法の有効性を確認している．

つぎに，人物関連の情報として，人物の住所などの位置情報を表す地名情報をWebページから抜き出し，地図上にマーク付けを行う手法を提案している．地名情報候補としては，住所と特定の建造物などのランドマークを挙げ，Webページから得られた文字列から形態素解析を用い，住所候補とランドマーク候補を抽出している．その後，氏名と住所，ランドマーク候補間の語間距離を用いて，住所，ランドマークを決定している．評価実験により，本手法の有効性を確かめている．

最後に，Web上で得られる人物に関する時系列事象の情報を，履歴書形式に変換する方法を検討している．履歴が表現されている時間と出来事を含む履歴文を抽出し，その中から，履歴書の内容として不適切な語，文を削除し，戸籍，学歴，経歴，受賞歴のカテゴリに分類し，履歴書に変換する方法を提案している．評価実験を行い，有効性を確かめている．

これらの成果は，Web 上から得られる人物情報に関する同姓同名人物問題の解消および人物関連情報の抽出と表示に新たな知見を提供し，情報検索分野、特に Web インテリジェンス分野の発展に寄与するところが大きい．よって，本論文の著者は，博士（工学）の学位を受ける資格を有するものと認める．