審 査 の 結 果 の 要 旨

氏　名　　ヴー　クァンミン

本論文は、「A Study on Name Disambiguation Using Web Directories（ウェブディレクトリを用いた人名の曖昧性解消に関する研究）」と題し、英文６章から構成されている。ウェブ検索において人名に関する文書の集合の中から同姓同名の人物を識別するための方法について論じたものである。人物を識別するために、ウェブディレクトリを援用してウェブ文書の中のトピックを見出し、トピックを表す単語の共起を評価することにより精度の高い同定手法を提案するとともに、従来手法との比較を実験的に行い、またデモシステムを構成して評価・検証している。

第１章は「Introduction」であり、研究の背景について論じている。インターネットとワールド・ワイド・ウェブ（WWW）空間の特徴を述べ、ウェブ空間から人物に関する情報を集めるニーズが高まっていることを指摘している。そして、本研究の位置づけと意義を述べ、提案手法の概要と特徴を概観している。

第２章は「Related Researches」であり、人名同定に関する既存の研究やその他のテキストマイニングに関する研究をまとめ、種々の手法を紹介している。特に本研究と関連する課題として単語の語義曖昧性解消の研究と人名曖昧性解消の研究をあげることができる。既存手法のアプローチとして、機械学習、ベクトル空間法、キーワード抽出手法、固有名詞抽出手法などに基づいた曖昧性解消方法を比較している。そして、既存研究と本研究の違いを説明した後、既存手法をウェブ文書に適用する際に問題となる点をまとめている。

第３章は「Using Web Directories to Improve Context Extractions for Name Disambiguations」と題し、本論文で新たに提案する手法の詳細を説明している。まず問題の困難さについて一般的に論じ、それに沿って典型的なベクトル空間法の欠点を述べている。これを克服するために、知識ベースを使う方法を提案し、このアプローチに基づき、インターネット上で利用できる種々のウェブディレクトリを知識ベースとして活用する方法を提案する。ウェブディレクトリはトピック毎に分類された文書集合である。様々なレベルの分野に分かれて編集されているこれらの文書集合は、文書量が多く、またトピックを表す単語を見つけやすいという特徴を持つ。このトピック単語を活用して曖昧性解消に使うという点が、本手法の本質的な特徴である。ウェブディレクトリの特徴ベクトルの計算式および文書特徴ベクトルの修正式の導出について議論した後、文書の修正特徴ベクトルを用いて、文書間の類似度を計算するアルゴリズムを詳説している。

第４章は「Extract Topics in Web Directories for Improvements of Name Disambiguations」と題し、ウェブディレクトリに含まれるトピックを用いて、同姓同名の曖昧性を解消する方法を提案している。トピックを抽出するための手法として、 Latent Dirichlet Allocation（LDA）を紹介し、これをウェブディレクトリに適用する方法を述べている。本来の LDA 手法をウェブディレクトリに適合させるための改善方法を提案する。次に、抽出したトピックを用いて、文書類似度の計算方法を提案する。抽出したトピックに基づき、単語のトピック特徴ベクトルと文書のトピック特徴ベクトルを計算する。これらの特徴ベクトルを用いて、元の文書の特徴ベクトルを修正し、最終的な類似度を計算するアルゴリズムを提案している。

第５章は「Experiments」であり、提案手法の有効性を検証するために行った実験を報告している。実験に使ったデータの用意について説明し、同姓同名の人物に関する文書とウェブディレクトリの準備方法を述

べている。同姓同名人物に関する文書として、Google 検索エンジンから検索した文書集合を使用し、ウェブディレクトリとしてGoogle、Dmozそして Yahooの三種のディレクトリを使用して評価実験を設計している。評価基準となるベースライン手法としてはベクトル空間法と固有名詞抽出手法を使っている。実験結果を分析し、ベースライン手法と比較しながら本手法の特徴と有効性を検証している。

第6章は「Name Disambiguation Demo System」であり、提案手法を用いて同姓同名の曖昧性を解消するデモシステムを紹介している。デモシステムは提案手法の内部の特徴を把握しやすいように作られたもので、ユーザの質問を Google 検索エンジンに転送し、検索結果をダウンロードする。そして、検索結果を並べ替えて、人名の曖昧性解消を行うというものであり、本手法を始めベースライン手法の相互比較を検証できるように工夫されている。

第7章は「Conclusions」であり、本研究の全体をまとめである。第一に、提案する手法の特徴と研究の成果を論じ、学術的な貢献についてまとめている。次に、同姓同名問題の拡張問題を議論して、提案した手法の適用できる範囲を示している。最後に、本論文で提案している知識ベースを使う提案手法を他のテキストマイニングに適用する可能性を議論し、全体としての結論を述べている。

以上これを要するに、本論文は、爆発的に増大する情報源からの人物に関するウェブ文書の検索に関して、同姓同名を識別して人物を同定する新しい手法として、ウェブディレクトリを活用して抽出した文書中のトピックを使用することにより同定精度を高めるという方式を提案するともに、既存手法と実験的な比較、およびデモシステムの構築によって、その有効性を実証し、その結果を情報検索やテキストマイニングの研究分野で有用な知見としてまとめたものであり、電子情報学上貢献するところが少なくない。

よって、本論文は博士（情報理工学）の学位請求論文として合格と認められる。