論文の内容の要旨

論文題目　　Web 上の文書における名前の持つ曖昧性解消に関する研究

氏名　小　野　　真　吾

Web上には多くの情報があり，近年では検索エンジンを用いることで，その情報を容易に得ることができる．しかし，たとえば人物を検索する場合を考えてみると，人名には同姓同名が存在するように曖昧性がある．そのため，人名をクエリとして検索エンジンに与えてWeb検索を行った際に，1つの名前についての検索結果が実際には異なる数人についての記述を集めてしまう場合がある．このような場合，検索を行うユーザは自分の探したい情報を検索結果の中から探す必要がある．また，この同姓同名の問題と似た構造を持つ問題として，“ACL”や“UT”などといった英数字からなる略称をクエリとした検索結果の曖昧性がある.これらを統一的な枠組みの問題として整理した．次に，我々はこの2種類の名前が持つ曖昧性について解消することを目的として研究を行った.

同姓同名の曖昧性解消の問題は，Webの拡大と検索エンジンの性能の向上により，この問題が広く知られるようになってきている．そのため，近年さかんに研究が行われており，2006年には初めてこの問題に対する評価型ワークショップ(Web People Search Task) が開催された．これにより，共通のデータセットも整備されつつある．一方，略称の問題に関しては，略称が出現する文書の中に出現する正式名称を推定する研究が従来から行われている．しかし，我々はこの問題が同姓同名の問題と類似の構造を持つ問題であり,同姓同名の問題の手法を適用できる可能性があると考えた.

我々はこれらの問題を,文書中に出現するクエリの文字列が参照する実体が同一である文書が同じクラスタに入るように文書クラスタリングを行う問題と考え,文書中の検索対象の人名に関連するソーシャルネットワーク情報についての類似度と,文書中の重要語情報を用いた類似度を用いたクラスタリングを行う手法を適用した．また，文書中の特定の情報にのみ注目してクラスタリング結果を再度クラスタリングする2段階クラスタリングを行う手法や,Web上から略称に関する知識を自動的に取得し，それを利用してWeb上の文書中の略称を分類する手法を提案した．さらに，同姓同名の曖昧性解消を行うシステムを実装し，実際にデモとして公開した．このシステ

ムは，同姓同名の人物を正しく分離する高いクラスタリング性能と，実用的な動作速度の両方の達成を目的としている．

本論文の構成は以下のとおりである．まず，第1章においては，序論として本論文の目的と位置付けを述べ,本論文で取り扱う同姓同名の曖昧性解消の問題や略称の曖昧性解消の問題について,問題の構造や特徴を説明する．第2章では，本論文に関連する研究について紹介している．

第3章では,名前が持つ曖昧性解消を行うための方法論を説明する.文書中から特徴語を抽出し,その情報を元に類似度計算を行う方法や，我々が試みたいくつかの文書クラスタリングの手法,また，我々が新たに提案した2段階クラスタリングや略称に関する知識の獲得と，それを利用した分類手法について述べる．第4章においては，前章において紹介した各手法を適用するための実装や，実際に開発したシステムについて述べる．実装については，Web上の文書を取り扱うための前処理や，実際に作成したシステム，また略称の曖昧性解消を行うための略所に関する知識の獲得について説明する．

第5章,第6章では,我々が行った実験とその結果について説明する.同姓同名の問題については,我々が独自に作成した日本語データセットおよび共通データセットであるWePSデータセットを,略称の問題については略称に関するデータセットを作成し実験を行った．さらに，我々が開発した同姓同名の曖昧性解消を行うシステムについて,速度を最も重視した際のクラスタリング性能についても実験を行った．性能評価の結果，これらの手法を問題ごとに適切に組み合わせることにより，同姓同名の問題，略称の問題のいずれにおいてもF-measureで0.75程度の性能を達成することができた．また，システムについて行われたアンケート結果によると，クラスタリング性能と動作速度の両方について，十分に目的が達成できているという評価が得られた．最後に，第7章で本論文の結論を述べる．