審 査 の 結 果 の 要 旨

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、これにともなって検索エンジンの結果上位に現れない、いわゆるロングテール部分の情報を入手しにくい問題が顕在化している。例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、ある名前についての検索結果には異なる数人について混在し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ことは難しい。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。

本論文はこのような名前が持つ曖昧性について解消することを目的として学位申請者が行ってきた研究をまとめたものである。上記のように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り研究が活発化した。この論文では、この問題をWeb 上の文書をクラスタリングする問題と捉え、文書中の検索対象の人名に関連する固有名の出現状態の類似度と、文書中に出現した重要語の類似度を用いたクラスタリングによって解決する手法を提案している。加えて、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を新規提案し、全体をまとめて名前の参照曖昧性解消システムとして実装し、実験評価を行った結果について述べている。

本論文は「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」と題し、上記の名前の参照曖昧性の解消方法およびシステム設計における課題を解決する方法に関して論じたものであり、下記に示す7章からなる。

第 1 章　序論では、本研究の背景、問題の定義と位置づけを行っている。

第 2 章　関連研究では、本研究の関連研究をまとめている。

第 3 章　同名の曖昧性解消の方法論は、本論文の主要な理論を説明している。すなわち、提案する名前の曖昧性解消アルゴリズムにおいて、使用する固有名抽出、重要語抽出について説明し、次にそれらを用いたクラスタリング手法を述べている。クラスタリングにおいては機械学習アルゴリズムであるディリクレ過程ユニグラム混合によるトピック推定などの適用を提案している。次に一度クラスタリングされた結果に対して、クラスタ内の重要語によって再度クラスタリングする手法を提案し、性能向上を図っている。

第 4 章　実装では、3 章に述べた方法を Web の検索エンジンから人名検索して得られた結果に適用し、数秒以内でクラスタリングを行う実装方法について述べている。

第 5 章　実験環境では、評価実験に用いるデータ、すなわち実際に収集した日本語の Web ページ集合からなるテキストコーパスおよび国際的なコンペティッション型タスクである WePS(Web People Search Task)の英語の Web ページ集合からなる評価用テキストコーパスについて述べている。

第6章　実験結果では、5章で説明したデータを用いて、提案した手法の実験評価を行った結果について述べ

ている。この結果、日本語のデータに関してはF-値で0.74、英語のWePS2006のデータでは、世界最高水準であるF-値で0.79を実現した。また、類似の構造を持つ略称の曖昧性解消についても実験し、F-値0.76という結果を得ている。

第 7 章結論は、本論文のまとめである。

以上を要するに、本論文は与えた名前で検索して得られたWebページの集合における、名前の参照曖昧性解消処理に関して、固有名、重要語を使う情報抽出、機械学習を用いたクラスタリング、1回目のクラスタリングを利用した再クラスタリングによって高い曖昧性解消能力を持つクラスタリング方法を提案した。また、提案したクラスタリング方法によるシステムの実装を行い、実装上の工夫によって高速な処理を実現した。また、実験評価によって高い性能を実証し、提案した手法の実用性を示した。これらの結果は情報理工学の発展に寄与するところが大きい。

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。