| | |
|---|---|
| 氏名 | 楠村幸貴（くすむらゆきたか） |
| 博士の専攻分野の名称 | 博士（工学） |
| 学位記番号 | 第 20418 号 |
| 学位授与年月日 | 平成18年3月24日 |
| 学位授与の要件 | 学位規則第4条第1項該当<br>基礎工学研究科システム創成専攻 |
| 学位論文名 | **不均一な文書を対象としたテキストマイニングに関する研究** |
| 論文審査委員 | （主査）<br>教授 西田 正吾<br>（副査）<br>教授 谷内田正彦　教授 佐藤 宏介 |

## 論文内容の要旨

本研究ではWeb上の不均一な文書集合からユーザの求める情報を整形して抽出し、まとめたものをユーザに提供するシステムの開発を目指すものである。このとき問題となる文書の不均一さには、語彙の不均一さ、記述形式の不均一さ、情報の構造の不均一さ、意味の不均一さの4つがある。

本研究はまず、語彙の不均一さと記述形式の不均一さに注目し、ネットオークションの商品の比較表を自動生成するシステムを開発した。このシステムでは、最小限の辞書と機械学習、Shallowな構文解析の技術を組み合わせて情報の収集・抽出を行う手法を提案し、このシステムが実際のユーザにとって有益な支援を行えることを示した。

次に、本研究ではこのような情報収集・抽出システムを構築する際の人的コストに焦点を当てた。実用的なシステムの設計には、システム構築時の人的コストとそれによって実現される精度・再現率のバランスが重要となる。この研究では、複数の情報収集・抽出方式に対しこのバランスを調査し、情報収集・抽出システムを構築する際の指針を明らかにした。

本研究では、語彙の不均一さと記述形式の不均一さに加え、情報の構造の不均一さの問題を解消するために、人的コストの調査結果を元に、辞書と学習を組み合わせた情報抽出手法を提案した。この手法では、HTML文書のDOM構造を解析して抽出ルールを構築し、抽出ルールを交叉させることで、従来の情報抽出手法よりも少ない人的コストで大量の情報を抽出できることを示した。

また本研究では、ネットオークションの評価コメントに存在する意味の不均一さを解消するシステムを開発した。このシステムでは、ネットオークションに存在する社会的な関係を利用し情報要約を行う手法を提案し、この手法が従来の要約手法よりも有効であることを示した。

これらの不均一さを扱うテキストマイニング技術により、今後のWebの発展に寄与できれば幸いである。

## 論文審査の結果の要旨

本論文はWeb上の不均一な文書集合からユーザの求める情報を整形して取り出し、比較表や要約を自動作成する情報収集・抽出システムの開発を目指したものである。本論文は、このシステムを開発する上で、Web上の文書に存

在する不均一さの問題に焦点を当てている。

本論文ではまず、ネットオークションの商品の比較表を自動生成するシステム（NTM-Agent）において、語彙の不均一さと記述形式の不均一さを解消する手法を提案した。これは最小限の辞書と機械学習、Shallow な構文解析の技術を組み合わせて商品の収集・抽出を行う手法である。本論文では実際のネットオークションに対して実験を行っており、この手法がある程度の精度を持ち、ネットオークション上の実ユーザに対して実際に支援可能であることを示している。

次に、語彙の不均一さと記述形式の不均一さを解消する枠組みとして情報収集・抽出システムを定義し、情報収集・抽出システムを構築する人的コストとそれによって実現される精度・再現率のバランスについて調査した。この調査では、e-マーケットプレースにおいて複数の情報収集・抽出方式を実装し、精度・再現率と知識の入力にかかる人的コストの関係を調べる比較実験を行い、個々の方式の性質を明らかにしている。さらにこれにより、情報収集・抽出システム開発時の設計方針を明らかにしている。

この知見を元に本論文では、辞書と学習を組み合わせたブートストラッピング・アルゴリズムを Web 上で効果的に用いる方法を提案した。これは HTML 文書の構造を解析する抽出テンプレートを生成することで構造化された記述形式から抽出を行い、抽出テンプレートを交叉させることで記述の多様性に対応するという手法である。これらの手法を組み込んだ抽出アルゴリズムは、数十個の語彙を元に数万個のレコードを高精度で抽出できた。このことより、本論文が不均一な Web 文書に対して高度な情報抽出技術を提供したと考えられる。

また、ネットオークションの出品者評価コメントを要約するシステムにおいては、意味の不均一さの問題を解消する手法として、ネットオークション上の社会的関係を用いて要約する手法（Social Summarization 法）を提案した。この手法はある書き手がまれにしか書かない記述を重要とすることで儀礼的な記述を削除する手法であり、従来の要約手法に比べて高精度でユーザにとって有益な要約を生成できることを示している。

以上のように、本論文は Web 上の情報を用いて何らかの意思決定を行うユーザに対して、Web した不均一な情報を統合することで高度な支援を提供する方法論に寄与するものであり、その実システムで確認されており、学術論文として価値あるものと認める。