【27】

氏名	池田和史（いけだかずし）
博士の専攻分野の名称	博士（情報科学）
学位記番号	第　２５８６４　号
学位授与年月日	平成25年３月25日
学位授与の要件	学位規則第４条第１項該当 情報科学研究科情報ネットワーク学専攻
学位論文名	Techniques for Analysis and Classification of Online Opinions （ネット意見活用のための解析および分類技術に関する研究）
論文審査委員	（主査） 教授　東野　輝夫 （副査） 教授　村田　正幸　　教授　村上　孝三　　教授　中野　博隆

論文内容の要旨

インターネットの普及により、Webページやブログ、Web掲示板などを通した一般ユーザによる情報発信が増加している。その中でもテキスト情報は発信の容易さなどから、主要な情報源となっている。これらのテキスト情報を解析、加工することでコンシューマおよびビジネス向けに情報を提示し、活用することの価値は極めて大きい。

本博士論文では、(1)くだけた表現を解析するための言語処理技術、(2)有害情報の高精度な検出技術、(3)情報発信者の属性情報の抽出技術、の3点をテーマに研究を行った成果について報告する。

一つ目の研究テーマでは、くだけた表現を含む文書であっても、通常の文書と同様に高精度に解析する手法を提案する。提案手法では、くだけた表現の修正候補文字列をくだけた表現の少ない文書から検索し、修正ルールを生成する。生成した多数の修正ルールから文脈に適した修正ルールを選択的に適用するために、検索結果における修正候補文字列の出現頻度、修正前後の文字列間における編集距離、修正前後の文の形態素解析結果の比較、を用いて修正ルールをスコアリングする手法を合わせて提案する。

二つ目の研究テーマでは、有害な情報を含む文書を高精度に検出するため、語の係り受け関係に基づいて高精度に有害文書を検出する手法と、WebページのHTMLを解析することで、テキスト情報のみからでは検出が困難である有害文書を検出する手法を提案する。語の係り受け関係に基づく手法では、既存の語の出現傾向に基づく有害文書検出手法において誤判定しやすい事例について、語を含む文節と係り受け関係にある文節の組を取り出し、有害性との関連を学習することで、誤判定を減少させ、検出精度を向上する。加えて、概念辞書を用いて語を含む文節と係り受け関係にある文節を抽象化することで、より多くの誤判定を訂正可能な手法を提案する。WebページのHTMLを解析する手法では、有害なWebページのHTMLに偏って特徴的に出現するような文字列を自動的に抽出する。抽出した各文字列の出現の傾向と有害性の関係は未知であるため、人工知能の識別器を用いて各文字列の出現傾向の特徴とWebページの有害性の関係を学習させることで、有害なWebページの検出を実現する。提案手法はWebページの本文の情報を利用しないため、既存の語の出現傾向に基づく手法で検出が困難なWebページも検出が可能である点が特徴である。このため、既存手法と組み合わせて利用することも有効である。

三つ目の研究テーマでは、情報発信者の年代や性別といった属性情報を抽出するため、情報発信者がこれまでに発信した情報に出現する語の特徴から発信者の属性を推定する手法を提案する。提案手法では、単純に語の出現傾向の特徴のみを利用する従来の方式に加えて、発信情報における文書間のリンク関係や発信者のコミュニティを分析することによって、発信者の属性を高精度に推定する。

本博士論文で取り扱う技術は、ネット意見を活用したサービスに差別化要素となる付加価値を与える重要な課題に挑戦するものである。性能評価実験を通して、提案手法は実用レベルの性能を達成することを確認した。また、これらの技術により実現される機能の一部はマーケティングツールとしてすでに実用化展開されている。

論文審査の結果の要旨

インターネットの普及により、Webページやブログ、Web掲示板などを通した一般ユーザによる情報発信が増加している。その中でもテキスト情報は発信の容易さなどから、主要な情報源となっている。これらのテキスト情報を解析、加工することでコンシューマおよびビジネス向けに情報を提示し、活用することの価値は極めて大きい。本博士論文では、(1)くだけた表現を解析するための言語処理技術、(2)有害情報の高精度な検出技術、(3)情報発信者の属性情報の抽出技術、の3点をテーマに研究を行った成果について報告している。

一つ目の研究テーマでは、くだけた表現を含む文書に対して、通常の文書と同様に高精度に解析する手法を提案している。提案手法では、くだけた表現の修正候補文字列をくだけた表現の少ない文書から検索し、修正ルールを生成している。生成した多数の修正ルールから文脈に適した修正ルールを選択的に適用するため、検索結果における修正候補文字列の出現頻度、修正前後の文字列間における編集距離、修正前後の文の形態素解析結果の比較、を用いて修正ルールをスコアリングする手法を合わせて提案している。

二つ目の研究テーマでは、有害な情報を含む文書を高精度に検出するため、語の係り受け関係に基づいて高精度に有害文書を検出する手法と、WebページのHTMLを解析することで、テキスト情報のみからでは検出が困難である有害文書を検出する手法を提案している。語の係り受け関係に基づく手法では、既存の語の出現傾向に基づく有害文書検出手法において誤判定しやすい事例について、語を含む文節と係り受け関係にある文節の組を取り出し、有害性との関連を学習することで、誤判定を減少させ、検出精度を向上させている。加えて、概念辞書を用いて語を含む文節と係り受け関係にある文節を抽象化することで、より多くの誤判定を訂正可能な手法を提案している。WebページのHTMLを解析する手法では、有害なWebページのHTMLに偏って特徴的に出現するような文字列を自動的に抽出している。抽出した各文字列の出現の傾向と有害性の関係は未知であるため、人工知能の識別器を用いて各文字列の出現傾向の特徴とWebページの有害性の関係を学習させることで、有害なWebページの検出を実現している。提案手法はWebページの本文の情報を利用しないため、既存の語の出現傾向に基づく手法で検出が困難なWebページも検出可能である点が特徴である。このため、既存手法と組み合わせて利用することも有効である。

三つ目の研究テーマでは、情報発信者の年代や性別といった属性情報を抽出するため、情報発信者がこれまでに発信した情報に出現する語の特徴から発信者の属性を推定する手法を提案している。提案手法では、単純に語の出現傾向の特徴のみを利用する従来の方式に加えて、発信情報における文書間のリンク関係や発信者のコミュニティを分析することによって、発信者の属性を高精度に推定している。

本博士論文で取り扱う技術はネット意見を活用したサービスに差別化要素となる付加価値を与える重要な課題に挑戦するものである。実環境での性能評価実験を通して実用レベルの性能を有していることを確認している。さらに、これらの技術により実現される機能の一部はマーケティングツールとしてすでに実用化されている。以上のような理由から、本論文は博士（情報科学）の学位論文として価値のあるものと認める。