2V-3

WEB上の相談事例とトラブルデータベースを利用した重要事案発見のための要因解析

八十岡　智章† 岡田　将吾† 新田　克己† 高橋　久尚††
本村　陽一††† 田中　智貴††††
東京工業大学大学院総合理工学研究科知能システム科学専攻† 統計数理研究所††
産業技術総合研究所††† 国民生活センター††††

1. はじめに

国民生活センターでは，消費生活に関するトラブル相談の受け付け，危害情報の収集，蓄積，これに基づいた情報提供などの業務を行っている．現在，大量の事例から危険事例を発見する作業は人手で行われており，今後は自動的に新規のトラブル事例を発見し警戒を出す機能が求められている．

一方で，WEB上にはトラブルデータベースに存在していない最新の相談事例など，WEB特有の相談事例が存在しているため，WEB上のデータを利用することは重大なトラブル事例の警戒のために有益であると考えられる．

以上の背景を踏まえ，本研究では警戒すべき危害情報などの重要事案の発見・検出を目指しトラブルデータベースとWEB上の相談事例データを利用して，危害事例の分類器の作成方法を提案・評価する．またWEB上の相談事例と過去に蓄積された事例から構築した分類器が，警戒すべき重要事案の発見にどのように寄与するかを考察する．

2. トラブルデータベースの概要

本研究で扱うトラブルデータベースは1事例に付き252項目からなる．2002～2011年度のデータが存在しており，総数は10,624,129件である．項目には，発生年月日，自然言語で記述されている件名・概要や，商品分類・契約先名（トラブルとなった相手の企業名）といった属性の他に，主観的に判断された事例の危険さを表す項目もある．本稿では危害(身体的危害の有無)・拡大損害の有無(商品が原因で身体または財産に拡大して損害が生じているか)の二つの項目に着目し，危険事例を早期に発見し警戒するために，危険事例の分類器の作成を行う．本研究で提案する分類器作成の手順を図1に示す．

Factor analysis for the detection of major issues using a trouble database and trouble case taken from the WEB
Tomoaki YASOOKA, Tokyo Institute of Technology

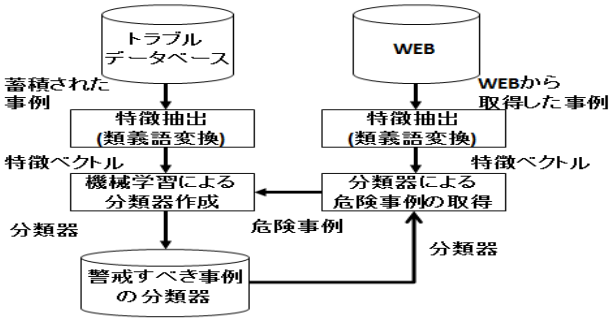

図1：分類器作成の手順

3. 提案する危害事例の分類アルゴリズム

3.1 類義語辞書とWeb上の相談事例の利用

類義語辞書には，アラジン・フォーラムが提供している負担・トラブル表現リストを用いる．リストには「風邪　病」のように，単語とその分類が記述されている．このリストは類義語辞書として扱うことが可能であり，本研究ではリストに存在する単語(風邪)を文章から抽出した場合，その分類名(病)に変換し，これを素性(特徴量)として利用する．

WEBデータにはヤフー株式会社が提供しているYahoo!知恵袋の投稿データを利用した．今回利用したデータは2004～2008年度の投稿のうち，投稿カテゴリ:「消費者問題」に属するデータとした．

3.2 基盤分類器の作成方法

まずトラブルデータベースから危害事例とタグ付けられたデータセットを正例，それ以外のデータセットを負例として用意する．

正例・負例すべてのデータに形態素解析を行い，件名・概要に含まれる内容語（名詞・動詞・形容詞）を抽出する．事例に内容語が含まれているか否かを1，0として数量化し特徴量として扱い，学習用データを用いて機械学習を行い分類器を作成する．

3.3 WEBデータを利用した分類器の作成方法

Yahoo!知恵袋の投稿データには危険事例のタグが付与されていないため，教師ラベル無しデータと見なせる．教師ラベル無しデータを学習に利用するために，[1]を参考にし，半教師付き学習手法の一つであるブートストラッピング法を用いた．以下に手順を示す

ステップ１：ラベル付きデータLを用いて分類器

を作成する．

ステップ２：ラベル無しデータU(WEBデータ)をステップ１で作成した分類器を用いて分類する．

ステップ３：ステップ２で分類された事例より分類器の出力値が閾値T以上のものを選び，Lに加える．

ステップ４：ステップ１からR回繰り返す．

4. 評価実験

トラブルデータベースに蓄積された危害事例とそれ以外の事例について各年度ごとにN件ずつ訓練データとして取り出す．テスト用データについても同様に危険事例と非危険事例をN件ずつランダムに抽出する．本研究の実験では，件数N = 2000の事例を用意し，各年度ごとに10回ずつ実験を行い，分類精度の平均値を算出した．また本研究では基盤分類器に線形SVMを用いた．

4.1 類義語辞書を利用した分類器の精度評価

本節では，類義語辞書の使用の効果を評価する．ここでは類義語変換処理を行った場合の線形SVMによる危険事例の分類精度と処理を行わなかった場合の分類精度の比較を図2に示す．

実験の結果，不要語削除と負担・トラブル表現リスト使用による分類精度の向上が見られた．また，拡大損害の有無に対する分類についても同様の結果が得られた．

4.2 WEBデータを利用した分類器の精度評価

実際の危険事例の分類・予測タスクを想定し，2008年度事例から作成した分類器で2009年度事例の危険事例判定を行ったところ，分類精度は92.5%となり，2009年度事例の分類器を使った時の94.9%よりも精度が落ちることがわかった．そこで，WEBデータを併用して学習することで，2009年度事例の分類精度を向上出来るか検証した．ラベル付きデータLをトラブルデータベースの2008年度事例とし，出力結果の閾値T = 2.0，繰り返し回数R = 1とした．その結果52件のWEBデータが危害事例と判定された．これをLに加え，再度学習を行い，2009年度事例の分類を行ったところ，精度は94.7%となり，WEBデータを併用することで分類精度を向上できることがわかった．

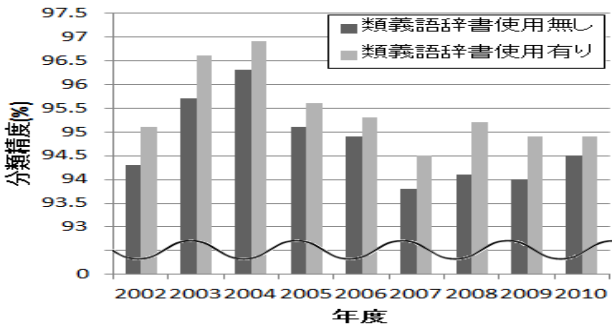

図2：危害事例の分類精度の比較
（トラブル表現リストの有無による比較）

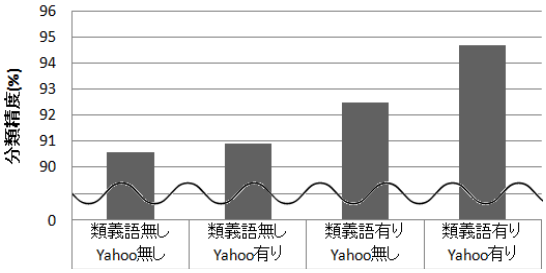

図3：危害事例の分類精度の比較
（類義語変換，Yahoo!知恵袋データ使用無しと有り）

2008年度事例で作成した分類器で2009年度事例を分類する際に，類義語変換・Yahoo!知恵袋データの併用する場合としない場合の分類精度の比較を図3に示す．

5. 重要事案発見の機能に関する考察

WEB上にはノイズとなる情報も多く存在しているため，提案手法においても閾値Tの値によってはWEB上の情報を追加することで分類精度を下げてしまう事態も起こりえる．Yahoo!知恵袋データから危険事例を取得する際，SVMの閾値T = 1.5では112件，T = 1.0では212件，T = 0.5では494件が取得され，取得した事例を併用して学習した．その後に分類を行った時の分類精度は閾値T = 1.5, 1.0では94.6%，T = 0.5では85.3%と精度が低下する．この結果閾値Tの設定が重要であることがわかる．

Yahoo!知恵袋から取得した危害事例の本文に目を通してみると，身体的被害に関する重要なトラブル事例であることが確認できた．また，分類器で誤分類された事例に目を通してみると，誤分類ではなく，ラベルの付け間違いである事例も発見された．この結果を踏まえ，提案手法で作成した分類器を，ラベルの付け間違い・付け忘れの確認，WEBからの重要事案発見のために使用出来ることが示唆された．

6. 結論

トラブルデータベースに蓄積された事例，類義語辞書，WEBデータを効果的に併用し，危害事例を分類する手法を提案した．評価実験の結果，提案手法は分類精度の向上に成功した．今後は[2]で提案されたアルゴリズムを用いた分類精度の向上，提案するシステム全体の実装，類似事例検索機能などの拡張を行う予定である．

参考文献

[1] 井上 裁都, 斎藤 博昭:ラベルなしデータの二段階分類とアンサンブル学習に基づく半教師あり日本語語義曖昧性解消, 自然言語処理18巻3号(2011)

[2] 小林 大祐, 松村 真宏, 石塚 満:知識検索サイトにおける不適切な投稿の分類，第21回人工知能学会全国大会(2007)