2S-1

大規模な共起辞書に基づく文書分類システムの試作

安藤 哲志 † 藤井 雄太郎 † 川口 将吾 ‡ 伊藤 孝行 †‡

† 名古屋工業大学大学院産業戦略工学専攻 ‡ 名古屋工業大学情報工学科

1 はじめに

近年，掲示板やブログといったユーザーが自由に投稿できるWebサイトが増加している．ユーザーが自由に投稿可能なWebサイトでは，未成年に有害な投稿がされることがあり問題となっている．多くのWebサイトでは，有害な記事が投稿された後に，人での確認によって対応を行なっている．しかし，人手による対応では運営コストが大きくなってしまう問題がある．

本稿では，有害な投稿記事を自動的に判定する手法の提案を行う．本稿での提案手法は，有害な文書である負例と無害な文書である正例から，共起情報を抽出した辞書を作成し，判定に用いる．本稿で共起は，単語A，BおよびCが同じ文書中に出現した場合に，単語A，BおよびCが共起した，としている．

2 関連研究

ベイジアンフィルタリング [1] はスパムメールフィルタリングに使われる手法であり，単語がスパムおよび非スパムに出現する回数を学習し，メールに含まれる特徴的な単語の結合確率を計算することでフィルタリングを行う．本稿で提案する手法は，このベイジアンフィルタリングを参考にしている．

サポートベクターマシンは (SVM) は精度の良い機械学習手法の1つとして知られている．SVMは学習するデータの特徴ベクトルから，データの分類をする超平面を求め，求めた超平面を用いて判定する手法である．SVMは選択するカーネル関数や特徴により，精度が大きく変わるという特徴がある．

Harmful Sentence Filtering System based on Large-scale Cooccurence Database

Satoshi Ando †, Yuutaro Fujii †, Shogo Kawaguchi ‡and Takayuki Ito ‡

†Techno-Business School , Nagoya Institute of Technology 466-8555, Nagoya, Japan

‡Dept. of Computer Science, Nagoya Institute of Technology 466-8555, Nagoya, Japan

{ando , fujii , kawaguchi }@itolab.mta.nitech.ac.jp , ito.takayuki@nitech.ac.jp

3 提案手法

3.1 提案手法の概要

本稿で提案する手法は，3つの単語の共起を用いて判定を行なっている．共起を用いた理由は文は多くの場合，主語，述語，目的語など複数の単語から構成されており，共起を用いることで文章の意味を考慮に入れることが出来るのではないかと考えたためである．また，掲示板などの文書は文の構成が正しいとは限らず，同じ意味の文でも単語の順序は前後することがあるため，提案手法は単語の出現順序は考慮に入れていない．

本稿で提案する手法は，以下の3ステップにより構成される．はじめに入力された文書を単語に分割する．ただし，分割された単語のうち，助詞や助動詞など単独では意味を成さない単語は単語から除く．次に分割された単語にブラックワードが含まれるかを確認する．ただし，ブラックワードは単独で有害であると判断できる単語である．ブラックワードが含まれている場合は有害な文書と判定し，含まれていない場合は次に進む．最後に分割された単語の共起の組み合わせと共起辞書から安全度を計算する．計算した安全度が閾値以下なら有害な文書，閾値以上なら無害な文書と判定する．

3.2 共起辞書の構築

本稿で構築した共起辞書について述べる．本稿では，単語 w_1，w_2 および w_3 の共起を (w_1, w_2, w_3) と記述する．共起辞書は，各正例および負例に含まれる，すべての共起を数え上げ，それぞれ合計したものから構成されている．共起辞書では，データサイズ圧縮のため，すべての単語をIDへとハッシュし，単語を整数型のIDとして取り扱っている．また，すべての共起は単語のIDによりソートを行っており，単語の順序は考慮に入れていない．また，共起辞書の構築は，単語をIDへとハッシュした後に分散処理ツールのHadoopを用いて構築している．本稿で作成した共起辞書のデータベース構造は，単語 w_1 のID，単語 w_2 のID，単語 w_3 のID，(w_1, w_2, w_3) の正例での出現回数，および (w_1, w_2, w_3) の負例での出現回数の5つの要素から構成され，それぞれ整数型で取り扱われている．

本稿では，正例および負例，それぞれ 10 万件から共起辞書の構築を行っており，正例および負例に含まれる単語の種類の総数は 108,675 件，および，構築した共起辞書の要素数は 1,498,732,588 件となった．ただし，正例は平均して約 20 単語を含み，負例は平均して 30 単語を含んでいる．

3.3 安全度の計算

本稿で提案する手法は，文書にブラックワードが含まれない場合，文書中の単語の共起から，文書がどれくらい安全であるかを示す安全度を求め判定を行う．安全度の計算方法について述べる．

共起 (w_1, w_2, w_3) の正例での出現確率 $P(w_1, w_2, w_3)$ を式（1）に示す．ただし，T_n は負例文書の総数，T_p は正例文書の総数，N_p は正例での共起 (w_1, w_2, w_3) が出現した回数，N_n は負例での共起 (w_1, w_2, w_3) が出現した回数であり，また，スムージングとして各出現回数に 1 を加算している．

$$P(w_1, w_2, w_3) = \frac{\{(N_p + 1)/T_p\}}{\{(N_p + 1)/T_p\} + \{(N_n + 1)/T_n\}} \quad (1)$$

式（1）をもとに，文書 S の安全度 $Safe(S)$ を式（2）で求める．ただし，$(w_1, w_2, w_3) \in S$ は S に含まれる共起の組み合わせであり，$P(w_1, w_2, w_3)$ を P と記述する．

$$Safe(S) = \frac{\prod_{(w_1,w_2,w_3)\in S} P}{\prod_{(w_1,w_2,w_3)\in S} P + \prod_{(w_1,w_2,w_3)\in S} (1 - P)} \quad (2)$$

4 評価実験

評価実験について述べる．評価実験では，テストデータの安全度を求め，判定を行う．本稿の実験ではテストデータに安全な文書を 14,064 件および有害な文書を 6,671 件を用いる．テストデータは 10 単語以上を含み，ブラックワードは取り除いたものを用いた．

実験環境について述べる．プログラミング言語に Ruby，形態素解析に Mecab，データベースに MySQL を利用し，分散データベースの構築に Hadoop を用いた．MySQL および Hadoop は 24GB のメモリをもつ計算機 6 台を用いた分散環境で利用している．

実験結果を図 1 に示す．ただし，グラフ中の横軸は安全度の範囲を示し，判定不能は文書に含まれる共起が共起辞書に存在せず，判定することができなかった文書であり，縦軸は安全度の範囲に含まれる件数である．

安全度の閾値を 0.5 とした場合の結果について述べる．無害な文書は 14,064 件中，13 件が判定不能であり，

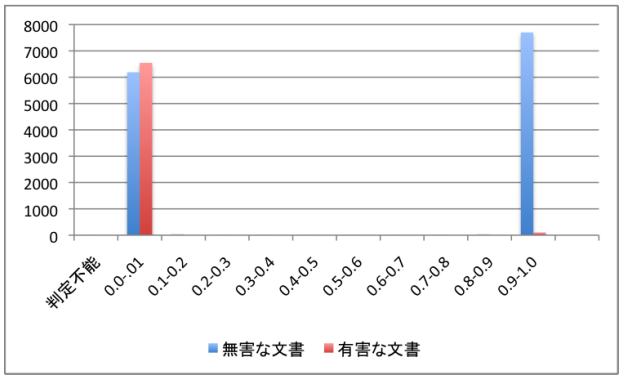

図 1: 実験結果

7,786 件が無害と判定され，6,278 件が有害と判定された．有害な文書は 6,671 件中，2 件が判定不能であり，6,558 件が有害と判定され，113 件が無害と判定された．無害な文書の再現率は 0.554，適合率は 0.985，および F 値は 0.709 であり，有害の文書の再現率は 0.983，適合率は 0.519，および F 値は 0.672 となった．

有害な文書は正しい判定をされる率が高く，無害な文書は正しい判定をされる率が低い結果となった．無害な文書の多くが有害と判定された理由は，負例に含まれる単語の平均数が正例に含まれる単語の平均数の約 1.5 倍あり，多くの共起が正例での出現回数より負例での出現回数が多くなってしまったために，判定結果が全体的に有害に傾いたのではないかと考えられる．

5 まとめと今後の課題

本稿では，有害文書を判定する手法として複数単語の共起情報を用いた手法の提案を行った．評価実験では，有害な文書は正しく判定される率が高く，無害な文書は正しく判定される率が低い結果となった．この結果は，正例より負例のほうが平均で約 1.5 倍の単語数を持っていたためであると考えられる．そのため，文書の総数だけではなく，正例および負例に含まれる単語の数を考慮に入れる必要があると考えられ，今後の課題としたい．

参考文献

[1] Paul Graham , ”A PLAN FOR SPAM” , http://www.paulgraham.com/spam.html

[2] 安藤哲志, 藤井雄太郎, 伊藤孝行,”有害文書判別のための多単語間共起情報辞書の構築とその応用”, 情報処理学会第 72 回全国大会,2010