3N-2

ベイズ分類器のスコアを素性に用いたSVMによる有害文書分類手法

藤井 雄太郎 †　　吉村 卓也 §　　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，ソーシャル・ネットワーキング・サービス (SNS) やブログ等の Web サイトが増加しており，未成年にとって悪影響を及ぼすような様々な情報が存在し，問題となっている．現在でもこれらの問題に対策を実施しているが，その多くは人の目視によるもので，時間的・金銭的コストの負担が大きくなってしまっている．そのため，効率良く有害な情報を適切に判別し，人や企業への負担を軽減するための研究が進められている．

本稿では，文章や画像などの様々な情報媒体の中でも特に文章に注目し，その中でも過度な性的描写を含む文章を対象とする．一般的に，SVM を用いた文書分類では，単語の出現頻度を素性としているが，本稿で提案する手法は，SVM の素性に共起を用いたベイズ分類器の Graham 方式と Robinson 方式によって計算した文書の有害確率を素性にし，分類を行う．

現在は，SVM を用いた文書分類に関する研究が盛んに行われている．Joachim ら [1] の研究では,SVM の素性として, 出現回数が 3 回以上の単語から，”or”，”and” 等のストップワードを除いた単語を対象としており，属性値として，単語の出現頻度を加味した IDF の値を採用している．

2 ベイズ分類器のスコアを素性に用いたSVMによる有害文書分類手法

本稿では，SVM の素性を作成するために，ベイズ分類器を作成する．ベイズ分類器では，WEB 上から収集した文書を用いて共起辞書を構築し，共起辞書を用いて文書の有害確率を計算する．

本稿における共起の定義を述べる．本稿では，1 つの文書に出現する全ての単語の組み合わせを共起と定義する．また，共起における組み合わせ爆発による計算量増加問題を回避するため，複数単語間の共起として，2 単語の共起関係を学習の対象とする．

共起辞書は，以下の流れで行う．また共起辞書の構築は MySQL で行う．

1. 学習データの収集

学習データの収集では，Web 上からクローラーを用いて学習データを自動で収集する．本稿では，2 ちゃんねる掲示板から学習データを収集する．

2. 学習データの振分け

学習データの振分けでは，収集した文書を無害文書の学習データである正例，有害文書の学習データである負例に分類する．分類では，まず，学習データを収集する段階で，ストップワード方式を用いて，有害文書のみを自動的に分類し，残りを人の目視によって分類する．今回ストップーワードは 200 種類程収集している．

3. 学習データの形態素解析による単語分割

学習データの形態素解析による単語分割では，学習データの文書を形態素解析器を用いて単語に分割する．本稿では，Mecab を形態素解析器に用いる．また，形態素解析器によって ’ 助詞 ’ 及び ’ 助動詞 ’ と解析された単語は，文書から取り除く．

4. 単語及び共起関係のハッシュ化

単語及び共起関係のハッシュ化では，分割した単語のハッシュ化を行う．ハッシュ化を行った後，学習データの単語を全て数字に置き換える．さらに, 学習データの 1 文書内で考えられる全ての単語の組み合わせを共起関係として，ハッシュ化する．ハッシュ化を行う事で，共起辞書のデータサイズを大幅に小さくできる．

5. 共起回数のカウント

共起回数のカウントでは, 共起関係をカウントし, データベースに格納する．

表 1 に，構築した共起辞書の内訳を示す．

表 1: 学習データの内訳

正例	100,000
負例	100,000
単語の種類	108,675
共起の種類	53,310,776

続いて，構築した共起辞書を用いて，文書の有害度

†Yutaro Fujii　†Atsushi Ando　†§Takayuki Ito
†Master course of Techno-Business Administration, Nagoya Institute of Technology
§Department of Computer Science and Engineering, Nagoya Institute of Technology

を確率的に求める．計算手法として，ベイジアンフィルタの Paul Graham 方式と Gray Robinson 方式の 2 つの手法を共起を用いて拡張する手法を用いる．

Paul Graham 方式では，[2] の手法での，単語の有害確率の代わりに，共起の有害確率を用いて，また正例，負例における単語の出現回数の代わりに，正例，負例における共起の出現回数を用いて分類を行う．Gray Robinson 方式では，吉村ら [3] の手法を用いる．

本稿では，作成したベイズ分類器を用いて素性の生成を行う．生成する素性は 2 つあり，1 つ目は Paul Graham 方式によって計算した文書の有害確率，2 つ目は Gray Robinson 方式によって計算した文書の有害確率を用いる．また，今回 SVM を用いるにあたり，LIBSVM をライブラリとして利用する．

3　実験結果と考察

本稿では，提案手法により無害文書，有害文書の各 8000 件のテストデータを用いて分類実験を行い，ベイジアンフィルタリングの Paul Graham 方式，Gray Robinson 方式及び，Joachim ら [1] の手法と比較実験を行う．テストデータは，有害文書は 2ch 掲示板の 18 禁カテゴリから，無害文書は 2ch 掲示板の 18 禁以外のカテゴリから収集した文書を用いる．また，本実験では，提案手法においてパラメーターチューニングを行った．カーネル関数はガウスカーネル，$\gamma = 2^{10}$ 及び $Cost = 2$ のパラメータにおける実験結果を示す．また，本実験において，精度を表す評価指標として F 値を採用する．図 1 に，テストデータにおける文書の有害確率分布を示し，図 2 に SVM による分類結果を示す．図の縦軸は G.Robinson による有害確率、横軸は P.Graham による有害確率を示し，赤色は有害文書，青色は無害文書の分布を示す．また，表 2 に，各手法の再現率，適合率及び F 値を示す．

表 2: 各手法における評価値

手法	適合率	再現率	F 値
提案手法	0.960	0.946	0.954
Paul Graham 方式	0.851	0.868	0.859
Gray Robinson 方式	0.951	0.826	0.884
Joachim らによる手法	0.813	0.675	0.738

表 2 を見ると，提案手法が全ての値において，最も高い値をとっている事がわかる．これは，Graham 方式の，文書の有害確率が 0.0 と 1.0 の両極端に分布する性質と，Robinson 方式の，文書の有害確率が平均的に 0.5 付近に分布する性質を組み合わせる事により，適切

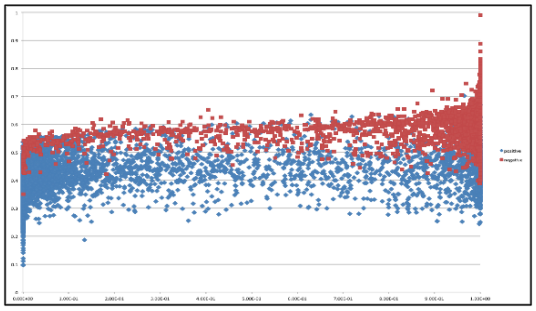

図 1: テストデータの有害確率の分布

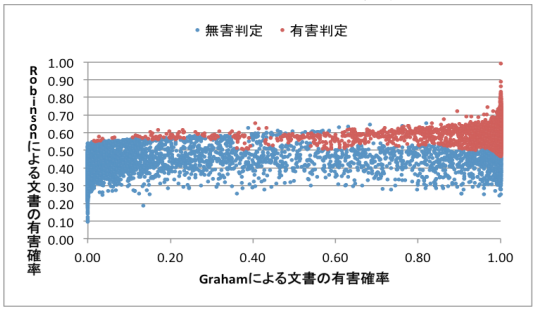

図 2: SVM による分類結果

な分類が行われたと考えられる．また計算時間に関して，Joachim らの手法との比較すると，本実験において 8000 件のテストデータを分類する際に，Joachim らの手法では，28734s かかるのに対して，提案手法は 39s という結果が得られた．これは，素性数の差が要因だと考えられ，Joachim 手法の素性数 12837 に対して，提案手法の素性数は 2 であり，SVM による学習及び分類に大きく影響したと考えられる．以上の事から，既存の SVM を用いた分類手法に比べ，提案手法は計算コストも抑えられたと言える．

4　まとめ

本稿ではベイズ分類器のスコアを素性に用いた SVM による有害文書分類手法の提案，及び他の手法との比較実験を行った．各手法において，提案手法が最も高い精度が見られた．

参考文献

[1] T Joachims, ”Text Categorization with Support Vector Machines:Learning with Many Relevant”, Machine Learning: ECML-98, 1998 - Springer

[2] Paul Graham: ”A Plan for Spam”, http://www.paulgraham.com/ spam.html

[3] 吉村卓也，藤井雄太郎，伊藤孝行，Robinson 型判定手法を用いた単語共起フィルタの検証， FIT2011， 2011