3D2-3

複数単語間の共起情報を用いた有害文章判定手法の提案

A Harmful Sentence Filtering Method based on cooccurence among Multiple words

安藤 哲志[*1] 藤井 雄太郎[*1] 伊藤 孝行[*1]
Satoshi Ando Yutaro Fujii Takayuki Ito

[*1]名古屋工業大学 産業戦略工学専攻
Techno-Business School , Nagoya Institute of Technology

In many sites, harmful writings are removed manually. Therefore, they use much time to deal with the harmful sentences. In this paper, we propose two methods to classify automatically harmful sentences and harmless sentences using the cooccourence relation among multiple words. Our methods enable to decrease the manual cost dramatically. In this paper, we show some experimental results. Accoding experimental results, we confirm the effectiveness of our methods.

1. はじめに

掲示板や SNS(Social Network Service) といったユーザーが自由に書き込めるサイトが増加している．ユーザーが自由に書き込みを行うことのできるサイトでは，未成年に有害な書き込みがされることがある．多くのサイトでは，有害な書き込みに対して対処を行っておらず，また，有害な書き込みへ対処を行っているサイトも多くは書き込みがなされてから人手による確認によって対処を行っている．しかし，人手による対処では，対処するためのコストや，対処するまでの時間が大きい．そこで本稿では，有害な書き込みを自動的に判定する手法の提案を行う．本稿で提案する手法は，有害な文書である負例と有害では無い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．

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，単語がスパムか非スパムどちらに特徴的に出現するかを学習し，メールに含まれる特徴的な単語の出現の割合を計算することでフィルタリングを行う．しかし，ベイジアンフィルタリングでは，単語の出現確率でフィルタリングを行うため，非スパムに特徴的な単語を多く含むスパムメールは非スパムメールであると判定してしまう問題点がある．

サポートベクターマシン (SVM)[2] は学習モデルの 1 つであり，機械学習の中でもっとも精度のよい手法の 1 つとして知られている．SVM は学習するデータから複数の特徴を抽出し，各特徴 (素性) から分類できる関数を求め，求めた関数によってデータを分類する手法である．SVM では選択するカーネル関数や素性により，精度が大きく変わるという特徴があり，カーネル関数や素性の選択が難しいという問題点がある．

連絡先: 愛知県名古屋市昭和区御器所町名古屋工業大学 19 号館 207,211 室
TEL:052-735-7968 FAX:052-735-7407
電 子 メ ー ル:{ando,fujii}@itolab.mta.nitech.ac.jp, ito.takayuki@nitech.ac.jp

3. 有害文書判定手法

本稿では，3 単語間の共起情報を用いた有害文書判定手法を提案する．提案手法は以下のグレイワードフィルタリング，ブラックワードフィルタリング，および，有害度の計算の 3 つのフィルタリングステップにより構成される．提案手法のフィルタリングの流れを以下に示す．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り，グレイワードはそれ単独では有害であるか有害で無いかが判断できない単語である．

1. 入力された文章を単語に分割する．単語を分割する際，いくつかの変形操作を行う．変形操作については後述する．
2. ブラックワードフィルタリング：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．ブラックワードが含まれている場合は有害な文書と判断し，含まれていない場合，次に進む．
3. グレイワードフィルタリング：分割された単語にグレイワードが含まれるかを確認する．グレイワードが含まれていない場合は有害では無い文書と判断し，含まれている場合は次に進む．
4. 有害度の計算：分割された単語の共起の組み合わせから有害度を計算する．計算した有害度が閾値以下なら有害な文書，閾値以上なら有害では無い文書と判定する．有害度の計算方法については後述する．

4. 有害度の計算

本稿で提案する有害度の計算について述べる．有害度の計算では共起情報を持つ共起辞書を作成し，判定に用いる．共起辞書は，有害な文書の集合である負例と，有害では無い文書の集合である正例から共起情報法を抽出し，構築する．

4.1 単語への分割

本稿では，形態素解析に Mecab[*1]を使用しており，単語を分割する際にいくつかの変形操作を行っている．変形操作を行っている単語にはたとえば”非線形”という単語があげられる．”非線形”を形態素解析した場合，”非”と”線形”に分割され，本来の意味と違った意味になってしまうため，結合操作を

[*1] http://mecab.sourceforge.net/

行っている．また，助詞や助動詞，副詞など単独では意味をなさない単語は形態素解析の結果からのぞいている．

4.2 共起辞書の作成

本稿で作成した共起辞書について述べる．本稿では，単語 w_1，単語 w_2，および単語 w_3 が同一文書にあり，単語間の出現場所が一定の距離 N 以内に出現した場合，単語 w_1，単語 w_2，および単語 w_3 が共起したものとする．単語がある一定の距離 N 以上離れている場合には単語同士の関係性が低いと考え，共起に含めていない．本稿では，関係性のある単語間の距離 N を 15 とした．提案手法では有害な文書である負例と有害で無い文書である正例からグレイワード g と他の単語 w_1，および w_2 の共起回数を抽出し，共起辞書を作成している．共起辞書は，グレイワード g，単語 w_1，w_2，正例での共起回数 $N_p(g, w_1, w_2)$，負例での共起回数 $N_n(g, w_1, w_2)$ から構成される．

共起辞書の作成に用いる元となるデータについて述べる．正例はブログおよび掲示板から収集したものから人手で有害では無いと判定した文書を使用し，負例は人手で収集した有害である文書および，クローラーによって収集したブログブラックワードが含まれている書き込みを使用した．グレイワードおよびブラックワードも人手での決定を行っている．本稿で作成した共起辞書の要素数，ならびに作成に使用した正例，負例，ブラックワードおよびグレイワードの数について表 1 に示す．

表 1: データベースの作成結果

説明	要素数
共起辞書	5,799,636
正例	5,196
負例	9,609
総単語数	79,109
ブラックワード数	180
グレイワード数	168

4.3 有害度の計算

有害度の求め方について述べる．提案手法での文書 $sentence$ の有害度 $HF(sentence)$ を以下の式 (1) で求める．ただし，$(g, w_1, w_2) \in sentence$ は文章 $sentence$ に含まれるグレイワード g と他の単語 w_1，および w_2 の共起の組み合わせであり，$P_p(g, w_1, w_2)$ は正例での g，w_1，および w_2 の共起の出現確率，つまり共起の出現回数 $N_p(g, w_1, w_2)$ を学習した正例の総数で割ったものであり，$P_n(g, w_1, w_2)$ は負例での g，w_1，および w_2 の共起の出現確率，つまり共起の出現回数 $N_p(g, w_1, w_2)$ を学習した負例の総数で割ったものである．

$$HF(sentence) = \underset{(g,w_1,w_2)\in sentence}{AVERAGE} \left\{ \frac{P_p(g, w_1, w_2)}{P_p(g, w_1, w_2) + P_n(g, w_1, w_2)} \right\} \quad (1)$$

5. 評価実験と考察

本稿では実験環境として，プログラミング言語に Ruby，データベースに MySQL，形態素解析ソフトに MeCab，および 24GB のメモリを持つ計算機を使用した．本稿では，評価実験として，有害な文書 100 件と有害では無い文書 100 件に対して有害度の計算を行った．本稿では，有害度の計算の際の閾値を 0.5 と設定した．実験で用いるデータはブログおよび掲示板から収集したグレイワードを少なくとも 1 単語含み，ブラックワードを含まない文書である．実験結果を図 1 に示す．ただし，図中の判定不能は評価対象の文書の共起が共起辞書に含まれておらず，有害度をつけることができなかったものである．

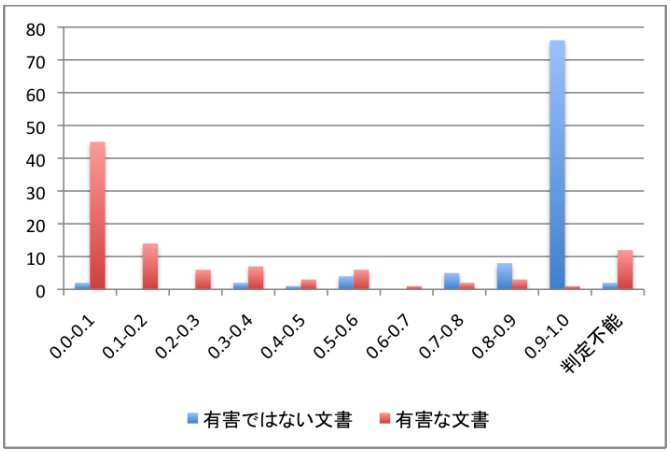

図 1: 実験結果

実験結果として，有害な文書 100 件中 75 件を有害な文書と判定し，13 件を判定することができず，有害では無い文書 100 件中 93 件を有害では無い文書と判定し，3 件を判定することができなかった．提案手法 1 での判定率は 84%となった．提案手法で誤判定をしてしまった文書は，共起辞書に含まれる共起が無く，判定することができなかった文書が多くあげられる．以上から，学習例は今回の用いた約 15,000 件では少ないということが考えられ，学習例を増やした共起辞書を作ることは，今後の課題である．

6. まとめと今後の課題

本稿では，ブラックワードフィルタリング，グレイワードフィルタリング，および有害度の計算の 3 つのフィルタリングステップからなる有害文書判定手法の提案を行った．有害度の計算では，有害文書である負例と有害で無い文書である正例から共起情報を抽出した共起辞書を作成し，共起確率を用いてフィルタリングを行った．評価実験の結果として，80%以上の精度を得ることができた．

今回提案した手法は，正例および負例から単語の共起関係を抽出し，利用しているため，判定したい文書の単語が正例および負例に全く含まれていない，あるいは単語の共起が共起辞書に含まれていない場合には判定することができない．そうした，新しい単語や組み合わせのみで構成される文書に対応することは今後の課題である．

参考文献

[1] A Plan for Spam , http://www.paulgraham.com/spam.html

[2] H. Drucker, C. Wu and V. Vapnik, "Support Vector Machines for Spam Categorization." IEEE Trans. On Neutral Networks, vol. 10, no 5, pp.1048-1054, 1999

[3] 津田裕一，八木秀樹，平澤茂一, "単語の共起を考慮に入れたナイーブベイズモデルによる文書分類", 第 29 回情報理論とその応用シンポジウム予稿集, pp.613-616,2006