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

安藤 哲志† 藤井 雄太郎‡ 伊藤 孝行†‡§*
†名古屋工業大学 産業戦略工学専攻 ‡名古屋工業大学 情報工学科
§マサチューセッツ工科大学 スローン経営大学院 *JST さきがけ 研究員

1 はじめに

SNS や掲示板のようなユーザーが自由に書き込みができるサイトが多くなっている．ユーザーが書き込めるサイトでは，未成年に有害な書き込みがなされることがある．多くのサイトではそうした書き込みに対処を行っていない．また，対処をしているサイトもほとんどは書き込みがなされてから人手により対処している．しかし，人手による対処では，コストや対処までの時間が大きくなってしまう．そこで，本稿では有害な書き込みを自動的に判別する手法を提案する．本稿で提案する手法は，有害な文書 (負例) と有害では無い文書 (正例) から共起関係を抽出し，判別に用いる．本稿では，単語 A，単語 B および単語 C が同じ文章中に出現した場合，単語 A，単語 B および単語 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，単語でフィルタリングをするため，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，精度のよい機械学習法として知られている．SVM は学習するデータの特徴から新しい次元を作成し，次元を増やすことで分類できる境界を求める手法である．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．

3 多単語間共起情報辞書の構築

3.1 概要

本稿で提案する判定手法は以下の手順で実行される．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り，グレイワードはそれ単独では有害であるか有害では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. 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. 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. 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．

3.2 有害度

本稿で提案する手法は，文章中にブラックワードが含まれず，グレイワードが含まれる場合，文章がどれくらい有害であるかを示す有害度を求めることで判定を行っている．文章 S の有害度 $HF(S)$ を求める式を式 (1) に示す．

$$HF(S) = \underset{(g,w_1,w_2)\in S}{AVERAGE}\left\{\frac{P(g,w_1,w_2)}{P(g,w_1,w_2)+N(g,w_1,w_2)}\right\} \quad (1)$$

ただし，$(g,w_1,w_2) \in S$ は文章 S に含まれるグレイワード g と他の単語 w_1，および w_2 の全組み合わせ，$P(g,w_1,w_2)$ は正例で g，w_1，および w_2 の共起が出現した回数，$N(g,w_1,w_2)$ は負例で g，w_1，および w_2 の共起が出現した回数である．

3.3 共起データベースの作成

本稿で作成した共起データベースについて述べる．本稿では，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と他の単語 w_1,w_2 の共起関係を抽出し，共起データベースを作成している．正例および負例はブログおよび掲示板から人手で収集を行い，

†Satoshi ANDO ‡Yuutaro Fujii †Takayuki ITO
†Techno-Business School , Nagoya Institute of Technology
‡Department of Computer Science, Nagoya Institute of Technology
§Sloan School of Management, Massachusetts Institute of Technology
* Researcher, PREST, Japan Science and Technology Agency (JST)

表1: 共起データベースの構造

説明	型
グレイワード g	varchar
単語 w_1	varchar
単語 w_2	varchar
正例での共起回数 $P(g, w_1, w_2)$	int
負例での共起回数 $N(g, w_1, w_2)$	int

表2: データベースの作成結果

説明	要素数
正例	5,296
負例	9,709
共起データベース	27,804,643
総単語数	56,805
ブラックワード数	250
グレイワード数	187

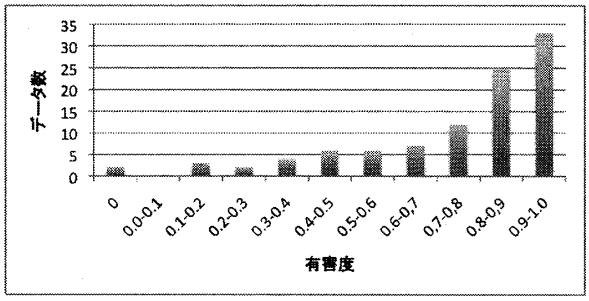

図1: 有害では無い文章の結果

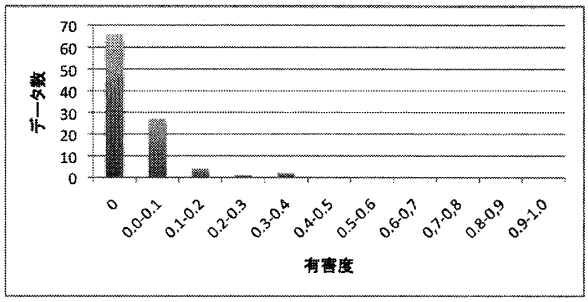

図2: 有害な文章の結果

グレイワードおよびブラックワードも人手で決定した．表1に作成した共起データベースの構造を示す．

また，本稿で作成したデータベースの要素数，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を表2に示す．

本稿では，形態素解析にMecab*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，結合操作を行っている．また，助詞や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 評価実験

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‡，および8GB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．実験で用いるデータはブログおよび掲示板から人手で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章である．実験結果を図1および図2に示す．本実験結果では，閾値を0.3とした場合，有害な文書は100件中98件が有害と判定され，有害では無い文書は100件中93件が有害で無いと判定された．誤判定をした文章には，共起データベースに含まれる共起がなく有害度が0になってしまったものなどがあげられる．

5 まとめと今後の課題

本稿では，共起関係を用いたフィルタリング手法の提案を行い，評価実験を行った．評価実験では，有害な文書および有害ではない文書でそれぞれ90％以上の精度で判別を行うことができた．ブラックワードやグレイワードを含まない文章を判定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] Peter D. Turney,"Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews",ACL 2002

*http://mecab.sourceforge.net/
†http://www.ruby-lang.org/ja/
‡http://www.mysql.com/