2 単語間の共起情報を利用した有害文章判別システム

藤井 雄太郎 †　　安藤 哲志 ‡　　伊藤 孝行 ‡§

†名古屋工業大学情報工学科　‡名古屋工業大学大学院産業戦略工学専攻　§MIT スローン経営大学院

1 はじめに

近年，携帯電話からの利用も可能となり，未成年ユーザが増加しているソーシャル・ネットワーキング・サービス (SNS) やブログ等では，未成年にとって悪影響を及ぼすような書き込みや画像，または動画を配信するユーザが存在し，問題となっている．そのため，現在では，効率良く有害な情報を適切に判別し，人への負担を軽減するための研究が進められている [1]．本稿では，配信される情報の中でも，文章に注目し，文章中の 2 単語間の共起情報を利用した有害文章判別システムを提案する．また，今回判別する文章の対象として，過度な性的描写を含む文章を対象とする．

2 関連研究

ベイジアンフィルタを使ったスパムメールを検出するシステムを構築した Graham ら [2] の研究が発表されてから，多くのシステムが開発されている．ベイジアンフィルタは，単純ベイズ分類器を応用し，対象となるデータを解析・学習し分類する為のフィルタである．ベイジアンフィルタをスパムメールに応用する場合，非スパムメールとスパムメールに出現する文字列に対する出現確率を学習し，その出現確率をもとに，ベイズ理論から新たに受信した電子メールに対して，スパムメールの検出を行う．スパムメールは，内容からも判定することは可能であるが，内容だけでなく，件名の書き方などそのスタイルが判定に有効であるといわれている．文字列の定義として，単語 (またはその語幹)，n 文字の連続する文字列などが用いられる．本稿の提案手法では，単語単体の出現確率ではなく，2 単語間の出現確率や距離を考慮する事で，より詳細な文章の情報を抽出する事で，精度の高いフィルタリングを目指す．

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

3 提案手法

3.1 辞書データベースの構築

まず，本稿における共起の定義として，文章中に出現したグレーワード gw の前後 20 単語以内の範囲に"単語"(cw_1,...,cw_n :(1≤n≤40)) が存在する時，cw_i と gw が共起関係 [$gw \Leftrightarrow cw_i$] にあると定義する．また，gw とは，単語の使用方法で有害な意味にもなり，無害な意味にも成り得る単語と定義し，"単語"は，動詞，名詞，形容詞，判別不能な品詞と定義する (以下，特定品詞)．

本稿では，有害文章判別を目的として，2 単語間の共起情報を元に辞書データベース (以下，辞書 DB) を構築した．辞書 DB は SNS 上に実在する多くの文章を用いる事で構築可能である．今回，辞書構築の元となる正例，負例に，yahoo ブログ*，goo ブログ†，2 ちゃんねる掲示板‡等の日記の文章や，掲示板の文章を用いた．形態素解析は Mecab を用いている．辞書の構築方法を以下に示す．1.gw を辞書に登録する．2. 収集した正例，負例から gw を検索する．3. 検索された gw から前後 20 単語以内にある特定品詞の単語 (cw_1,...,cw_n) を抽出する．4.[$gw \Leftrightarrow cw_i$] の出現回数をそれぞれカウントし，[$gw \Leftrightarrow cw_i$] 間の距離 $l(gw,cw_i)$ 毎にカウントをデータベースに登録する．ただし，ブラックワード bw はその単語単体で有害な意味になる単語と定義する．表 1 に辞書 DB の構造を示す．

表 1: 辞書 DB の構造

Field	説明
black_word	ブラックワード (bw)
gray_word	グレーワード (gw)
cooccur_word	gw と共起して出現した単語 cw_i
dist_5_p	$l(gw,cw_i)$≤5 の出現回数 (無害文章)
dist_10_p	6≤$l(gw,cw_i)$≤10 の出現回数 (無害文章)
...	...
dist_20_p	16≤$l(gw,cw_i)$≤20 の出現回数 (無害文章)
dist_5_n	$l(gw,cw_i)$≤5 の出現回数 (有害文章)
...	...
dist_20_p	16≤$l(gw,cw_i)$≤20 の出現回数 (有害文章)

*http://blogs.yahoo.co.jp/
†http://www.goo.ne.jp/
‡http://www.2ch.net/

3.2 有害文章判別アルゴリズム

試作した有害文章判別システムのアルゴリズムについて述べる．有害文章の判別は以下の方法で行う．

1. ユーザからの入力文 *text* を形態素解析し，単語に分割する．
2. 分割した単語から特定品詞を抽出する．
3. 抽出した単語に *bw*, 及び *gw* が含まれているかを調べる．
4. 3で調べた以下のパターン(1),(2), 及び(3)によって文章を判別する．(1) *bw* が含まれている場合の場合，*text* を有害な文章と判別する．(2)*bw*，及び *gw* 共に含まれていない場合 の場合，*text* を無害な文章と判別する．(3) *gw* のみが含まれている場合の場合，5を行う．
5. 2単語間の共起情報によって構築した辞書を用いて，入力文 *text* の安全度数 $S(text)$ を計算する．

$S(text)$ の計算方法は，*text* に出現する *gw* の前後20以内に存在する特定品詞の単語 $(cw_1,...,cw_n)$ を抽出し，*gw* と cw_i の単語間の距離 $l(gw,cw_i)$ を求める．続いて，単語間の距離 $l(gw,cw_i)$ によって辞書DBから単語 cw_i の安全度数 s_i を求める．また，*dist_l_p* は上記の表1の要素を表す．式(1)に計算式を示す．

・$l(gw,cw_i) \leq 5$ の時

$$s_i = \frac{(dist\_5\_p) * 2 + \sum_{10} dist\_l\_p}{(dist\_5\_p) * 2 + \sum_{10} dist\_l\_p + (dist\_5\_n) * 2 + \sum_{10} dist\_l\_n} \quad (1)$$

$$(l = 5, 10, 15, 20)$$

以下，同様に $l(gw,cw_i)$ によって辞書DBからの情報に重みをつけ，全ての単語 $(cw_1,...cw_n)$ に対して s_i を計算する．最後に，式(2)で，s_i の平均を計算し，その値を $S(text)$ とする．

$$S(text) = \frac{\sum_{i=1}^{n} s_i}{n} \quad (2)$$

6. 事前に設定した閾値 T と $S(text)$ を比較して，閾値以下ならば，*text* を有害な文章と判別する．

本稿における閾値の設定は，辞書構築時に収集した負例からランダムに文章を50個抜き出し，抜き出した負例以外の負例と正例で辞書を再構築し，それらの50個の文章の安全度数を計算する．これをさらに50回繰り返し，2500個の有害文章の安全度数の平均を閾値とする．今回は，上記の実験から，閾値 T=0.1 とした．

4 実験結果と考察

本章では yahoo 知恵袋*から取得した無害なテストデータ100個と有害なテストデータ100個を用いて，有害文章の判別実験を行う．実験方法は，yahoo 知恵袋の「アダルト」カテゴリから有害テストデータを取得し，それ以外のカテゴリから無害テストデータを取得した．それぞれのテストデータの安全度数 $S(text)$ を計算し，事前に設定した閾値と比較する事で，有害な文章かの判別を行い，システム全体の判別率 R((正判別した文章数/判別を行った文章数)*100) を計算し，精度を明らかにする．今回の実験では，$S(text)$ は小数点第2位を四捨五入した値とする．また，*text* から *bw* を検出した際には-1の値を，*bw*, 及び *gw* 共に検出されなかった場合には2の値を $S(text)$ とする．

*http://chiebukuro.yahoo.co.jp/

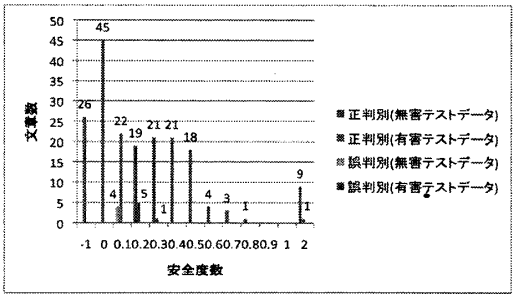

図1: 安全度数別の文章数の分布と判別結果

図1に実験結果として，安全度数別の文章数の分布と判別結果を示す．誤判別した無害テストデータは4個，正判別した無害テストデータは96個となった．これより，無害テストデータの判別率 R_p=96%という結果になった．続いて，誤判別した有害テストデータは7個，正判別した有害テストデータは93個となった．これより，有害テストデータの判別率は R_n=93%という結果になった．これらの結果から，R=94.5%となった．以上より，本システムでは90%以上の精度で有害文章を判別する事がわかった．

5 まとめ

本稿では2単語間の共起情報を利用した有害文章判別手法の提案，及び実在するSNSの文章を用いた評価実験を行った．評価実験では，9割以上の精度で正しい判別が可能である事がわかった．

参考文献

[1] 小林大祐，松村真宏，石塚満，"知識サイトにおける有害情報のフィルタリング知識の表出化"，第20回人口知能学会全国大会 2006

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