5S-3

複数単語間の距離情報及び共起情報を利用した文書分類手法の提案

藤井 雄太郎 †　　安藤 哲志　†　　伊藤 孝行 †§
† 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，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，文章中の 2 単語間の共起情報と距離を利用し，グレーワードという概念を用いた有害文書分類手法を提案し評価実験を行い，ベイジアンフィルタとの比較も行う．また，今回分類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 [1] 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．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．

3　提案手法

3.1　グレーワードの定義

本研究では，グレーワードという概念を用いる．グレーワードとは単語の意味が無害にも有害にも成り得る単語と定義する．例を挙げると，麻薬と速さを表す「スピード」という単語の事を言う．グレーワードは共起による計算量を減らす事ができ，比喩表現が多く用いられる性的描写の分類に有効であると考えられる．今回は、３人の学生で多数決をとり，多数の単語の中からグレーワードを選択した．

3.2　辞書データベースの構築

本稿では，動詞，名詞，形容詞，判別不能な品詞を単語として利用する (以下，特定品詞)．さらに，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．今回，有害文章判別を目的として，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する．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．表 1 に辞書 DB の構造を示す．

表 1: 辞書 DB の構造

説明	データ数
グレーワード (gw)	300
gw と共起して出現した単語 cw_i	1271547
$l(gw,cw_i)$≤5 の出現回数 (無害文章)	1438953
6≤$l(gw,cw_i)$≤10 の出現回数 (無害文章)	1271547
...	...
16≤$l(gw,cw_i)$≤20 の出現回数 (無害文章)	884029
$l(gw,cw_i)$≤5 の出現回数 (有害文章)	1737576
...	...
16≤$l(gw,cw_i)$≤20 の出現回数 (有害文章)	1390284

3.3　有害文書分類アルゴリズム

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．

†Yutaro Fujii　†Atsushi Ando　†§Takayuki Ito
†Master course of Techno-Business Administration, Nagoya Institute of Technology
§Sloan School of Management,Massachusetts Institute of Technology

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 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) によって文章を判別する．(1)gw が含まれていない場合，$text$ は分類不能とする．(2) gw が含まれている場合， 5 を行う．
5.. 辞書 DB を用いて，入力文 $text$ の安全度 $S(text)$ 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 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

$$s_i = \frac{(dist\_5\_p) * 3 + \sum_{10} dist\_l\_p}{(dist\_5\_p) * 3 + \sum_{10} dist\_l\_p + (dist\_5\_n) * 3 + \sum_{10} dist\_l\_n} \quad (1)$$

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

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，$text$ を有害な文章に分類する．

4 実験結果と考察

本章では無害な文章 100 件と有害ナ文章 100 個をテストデータに用いて，分類実験を行う．テストデータにはブラックワードを含まないものを採用した．実験方法はそれぞれの文章の安全度 $S(text)$ を計算し，閾値と比較する事で有害な文章の分類を行い，精度を明らかにする．また今回の実験での閾値は 0.2 とする．

図 1 に実験結果として，安全度数別の文章数の分布と結果を示す．無害な文章では 69 件が無害に，22 件が有害と分類された．9 件は分類不能となり，無害な文章の 69%が正しく分類された．有害な文では 3 件が有害に，4 件が無害に分類された．8 件が分類不能となり，有害な文章の 83%が正しく分類された．分類不能となったのは，文章中にグレーワードが出現しなかったためである．全体的に安全度が 0 に近い値を示す傾向にあり，これは負例と正例の学習量のバランスがとれていない事が原因であると考えられる．

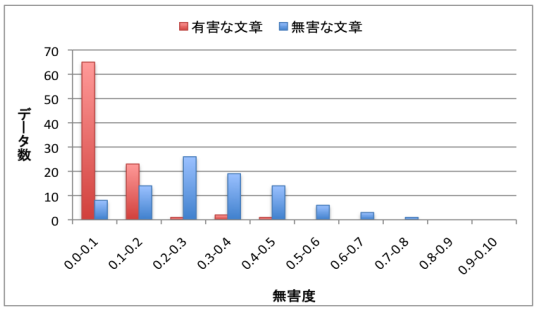

図 1: 安全度別の文章数の分布と判別結果

5 ベイジアンフィルタとの比較

今回，上記にもあったベイジアンフィルタと本研究で提案する手法との比較を行う．図 2 に 4 章で行った評価実験と同条件で行った実験結果を示す．ただし，安全度の基準が異なるため，閾値は最適である log(-1.0E20) とする．

実験では，無害な文章 100 件中，74 件が無害に分類され，26 件が有害に分類された．これより無害な文章の 74%が正しく分類された．有害文章 100 件中，62 件が有害に分類され，38 件が無害に分類された．これより，有害文章の 62%が正しく分類された．ベイジアンフィルタと比較すると，有害な文章に対する精度はわずかに勝っているが，その文無害な文章の精度が下がってしまった．

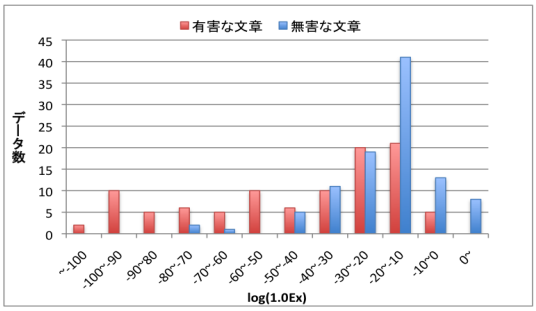

図 2: ベイジアンフィルタの文章数の分布と判別結果

6 まとめ

本稿では複数単語間の共起情報及び距離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．ベイジアンフィルタとの比較では，わずかだが良い結果が得られた．

参考文献

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