3D2-4

有害情報フィルタリングのための 2単語間の距離及び共起情報による文章分類手法の提案

Developing a method based on 2-word co-occurence information for filter harmful information

藤井　雄太郎[*1]　安藤　哲志[*1]　伊藤　孝行[*1*2]
Yutaro Fujii　Satoshi Ando　Takayuki Ito

[*1]名古屋工業大学大学院産業戦略工学専攻
Master course of Techno-Business Administration, Nagoya Institute of Technology

[*2]マサチューセッツ工科大学
Massachusetts Institute of Technology

Recently, social networking services, e.g., Mixi, Facebook, MySpace, etc., have been gathered much attention. However, there is a problem that some intended users upload harmful information for people into the services. This paper aims at developing a system that filter harmful information for people based on 2-word co-occurence information. In the experiment, we show the correctness of filtering in our system by using the real data.

1. はじめに

近年，携帯電話からの利用も可能となり，ソーシャル・ネットワーキング・サービス (SNS) やブログ等の発達に伴い，未成年ユーザの数も増加している．しかし，SNS 上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．現在は，人による情報の確認や巡回，ユーザによる通報などの対策が多くとられている．しかし，情報量が莫大な SNS では，発信される情報に対して上記のような対策では対応しきれず，多くの人的コストや時間がかかってしまう．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．

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章では，２単語間の距離及び共起情報による文章分類手法について述べる．３章では，本研究で行った評価実験について述べる．４章では，関連研究について述べる．５章では，本稿をまとめた後，今後の課題について述べる．

2. 2単語間の距離及び共起情報による文章分類手法

本章では，3.1 節で，本章の概要を述べる．3.2 節では本研究で用いた辞書データベースの構築方法及び辞書データベースの内訳に関して述べる．3.3 節では，学習データの収集方法について述べる．3.4 節では，２単語間の距離及び共起情報による文章分類手法のアルゴリズムについて述べる．

2.1 概要

本研究で提案する文章分類手法では，Web 上の学習データを収集し，２単語間の共起情報を抽出して辞書データベースを構築した．学習データの収集方法は，主に手作業で行った．文章分類アルゴリズムは辞書データベースから２単語間の距離及び共起情報を用いて，文章の安全度数を計算し，閾値と比較する事で行う．

連絡先: 藤井雄太郎，名古屋工業大学大学院産業戦略工学専攻 伊藤孝行研究室 〒 466-8555　愛知県名古屋市昭和区御器所町 19 号館 207 号室 fujii@capecod.mta.nitech.ac.jp

ここで，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”列が存在し，$\{(cw_1,...,cw_i,...,cw_n) : (1 \leq n \leq 40)\}$ と表す時，cw_i と *gw* が共起関係 $c(gw,cw_i)$ 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)．*gw* を定義した理由は，性的描写には一般的な表現が比喩的に用いられる事が多く，無害な一般的表現と有害な一般的表現の判別を行うためである．

2.2 辞書データベースの構築

本稿では，有害文章分類を目的として，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．形態素解析は Mecab を用いる．辞書の構築方法を以下に示す．

1.*gw* を辞書に登録する．
2.*gw* が含まれる文章を Web 上から収集する．
3. 収集した文章を人手で正例，負例に分別する．
4. 収集した文章中の *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 Mecab を用いて抽出する．
5.$c(gw,cw_i)$ の出現回数をそれぞれカウントし，$c(gw,cw_i)$ 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．ここで，共起単語間の距離 $l(gw,cw_i)$ とは，$c(gw,cw_i)$ 間に含まれる特定品詞の単語の数と定義する．

表 1 に辞書 DB の構造を示す．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．ここで，dist_5 とは，$c(gw,cw_i)$ において，*gw* から cw_i までの単語の距離が１以上５未満で共起している単語の組み合わせの出現回数をカウントしたもので，dist_10 は単語間の距離が 6 以上 10 未満で共起している単語の組み合わせの共起回数をカウントしたものである．dist_15 と dist_20 についても同様である．

2.3 データ収集方法

本研究では，有害文章の分類を行うために，学習データの収集を行った．収集する学習データは *gw* を含む文章とした．学習データは yahoo ブログ[**]，goo ブログ[*†]，２ちゃんねる掲

[**]http://blogs.yahoo.co.jp/
[*†]http://www.goo.ne.jp/

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	$1 \leq l(gw,cw_i) \leq 5$ での c(gw,cw_i) の出現回数
dist_10	$6 \leq l(gw,cw_i) \leq 10$ での c(gw,cw_i) の出現回数
dist_15	$11 \leq l(gw,cw_i) \leq 15$ での c(gw,cw_i) の出現回数
dist_20	$16 \leq l(gw,cw_i) \leq 20$ での c(gw,cw_i) の出現回数

示板[**]等の日記や掲示板の文章を収集した．データの収集方法は，正例と負例で異なる．正例は上記の Web サイトから手作業により，収集した．負例は 2 ちゃんねる掲示板の成人専用のスレッドからクローラーにより自動収集を行った　．正例の収集を手作業による収集方法で行った理由として，有害な文章が存在しない信頼できるサイト (例えば，ニュースサイト等) の数が少なく，さらにそれらのサイト内の文章自体にも *gw* を含む文章が少ないため，自動収集では多くの正例を収集できなかったためである．表 2 に辞書 DB の内訳を示す．自動収集が可能な分，負例の文章数の方が正例の文章数よりも多くなっている．

表 2: 辞書 DB の内訳

データの種類	データ数
ブラックワード	249
グレーワード	187
正例	5305
負例	9079
共起の組み合わせ	8156156

2.4 有害文章分類アルゴリズム

提案する有害文章分類アルゴリズムについて述べる．有害文章の分類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) によって文章を分類する．
(1) *bw* が含まれている場合の場合，*text* を有害文章に分類．
(2)*bw*，及び *gw* 共に含まれていない場合，*text* を無害文章に分類．
(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．
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* を有害文章に分類．

$S(text)$ 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 c(gw,cw_i) の安全度数 s_i を求める．また，count[gw,cw_i,l,(p or n)] は，c(gw,cw_i) の距離が (l-4) 以上 l 以下で，正例 (p) または負例 (n) として出現した回数を表す.P_i は正例において c(gw,cw_i) の出現回数に $l(gw,cw_i)$ によって重み付け (*2) を行った値とし，式 (1) に計算式を示す．また，同様に N_i は負例において c(gw,cw_i) の出現回数に $l(gw,cw_i)$ によって重み付けを行った値とし，式 (2) に計算式を示す．安全度数 s_i を求める計算式を式 (3) に示す.

・$l(gw,cw_i) \leq 5$ の時

$$P_i = count[gw, cw_i, 5, p] * 2 + \sum_{l=10} count[gw, cw_i, l, p] \qquad (1)$$

$$N_i = count[gw, cw_i, 5, n] * 2 + \sum_{l=10} count[gw, cw_i, l, n] \qquad (2)$$

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

$$s_i = P_i/(P_i + N_i) \qquad (3)$$

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，*gw* と全ての単語 (cw_1,...cw_n) の c(gw,cw_i) に対して s_i を計算する．最後に，式 (4) で，s_i の平均を計算し，その値を $S(text)$ とする．

$$S(text) = \sum_{i=1}^{n} s_i/n \qquad (4)$$

本稿における閾値は，負例からランダムに文章を 50 個抜き出し，辞書を再構築し，それらの文章の安全度数を計算する．これを繰り返し，2500 個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．

3. 評価実験

本章では，提案手法での文章分類精度を明らかにするための性能実験と，既存の文章分類手法との比較を行う比較実験について述べる．性能実験においては，yahoo 知恵袋[*†]から取得した *gw* を含む文章を対象と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，有害文章分類の実験を行い，文章分類の精度を明らかにする．比較実験では，上記と同様のテストデータを用いて，提案手法とベイジアンフィルタとの比較を行い，有害文章と無害文章の安全度数の差異化が図れているかどうかの視点から提案手法の有効性を示す．

3.1 性能実験と結果

性能実験では，Web から収集した実験データを用いて，それらの安全度数を計算し，閾値と比較する事で，文章の判別を行う．実験データは，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，有害文章分類の精度を明らかにする．無害文章における文章分類精度 R_p((正判別した無害テストデータの数/無害テストデー

[**]http://www.2ch.net/

[*†]http://chiebukuro.yahoo.co.jp/

タ数)*100)，有害文章における文章分類精度 R_n((正判別した有害テストデータの数/有害テストデータ数)*100) を計算し，精度を明らかにする．また，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して，$text$ 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．

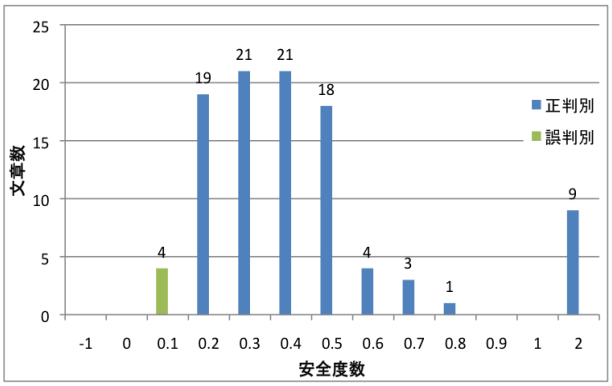

図 1: 無害文章分類における安全度数別の文章数の分布と判別結果

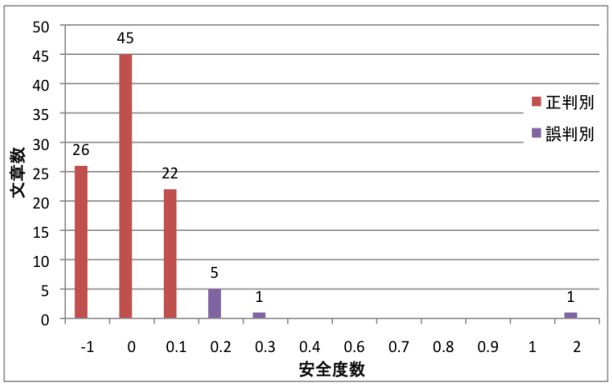

図 2: 有害文章分類における安全度数別の文章数の分布と判別結果

図 2 に実験結果として，無害テストデータの分類における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，R_p=96%という結果になった．続いて，図 3 に有害テストデータの分類における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．以上より，本システムでは無害文章，有害文章ともに 90%以上の精度で判別する事がわかった．

3.2 比較実験と結果

比較実験では，性能実験で用いたテストデータと同様の実験データを用いて，提案手法とベイジアンフィルタの内容でそれぞれ文章の安全度数を計算する．また，今回の実験でベイジアンフィルタに用いる学習データは，本研究で用いた辞書データベースを構築する時の学習データを用いている．よって，学習データは提案手法とベイジアンフィルタとで同様のデータとなっている．比較実験では，どれだけ文章の安全度数の分布が無害文章と有害文章で異なっているかを比較し，評価する．

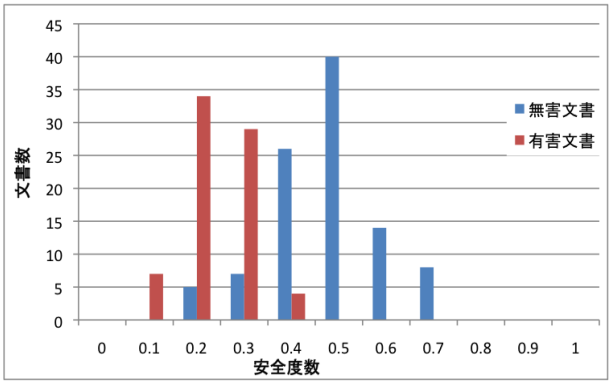

図 3: ベイジアンフィルタにおける安全度数別の文章数の分布

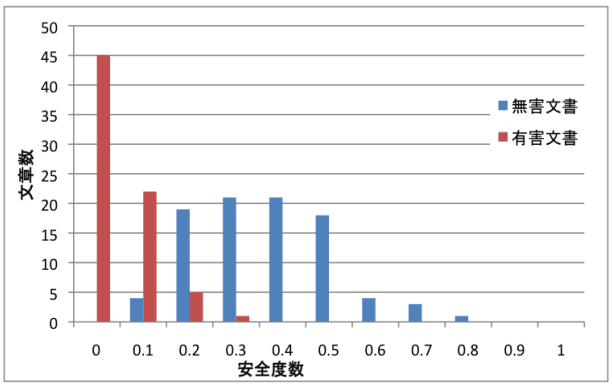

図 4: 提案手法における安全度数別の文章数の分布

図 3 に実験結果として，ベイジアンフィルタによる安全度数別の文章数の分布を示す．図 3 から，無害文章は 0.4，0.5 付近の数値に集中して分布しており，有害文章は 0.2，0.3 付近の数値に集中して分布している．全体的に見ると，無害文章と有害文章共に中央に集中している事がわかる．続いて，図 4 に提案手法による安全度数別の文章数の分布を示す．図 4 から，無害文章の安全度数は 0.2～0.5 までの数値に集中しており，ベイジアンフィルタに比べ，0.2，0.3 付近の値が大きくなってしまっているが，有害文章の安全度数は 0，0.1 付近に集中し，極めて 0 に近い値となっている．無害文章が 0.2～0.5 付近に分布している理由として，正例と負例のデータ量の偏りによるものだと考えられる．２つの手法を比較すると，無害文章の分布に対して，わずかに提案手法の方が，有害文章の安全度数に偏りができている．この事から，提案手法がベイジアンフィルタよりも，有害文章と無害文章の特徴を抽出できてると考えられる．

3.3 閾値の考察

本研究においては，閾値の設定方法の 1 つとして，辞書の解析結果を用いた．本システムの判別精度は閾値に依存し，ユーザは環境に応じた閾値を設定する必要がある．そこで，本章では閾値の変化におけるシステムの精度を，再現率と適合率を算出し，考察を行う．図 4 に 4.2 章で行った評価実験で用いたテストデータから有害文章を検索する際の適合率，及び再現率と閾値の関係を表すグラフを示す．

このグラフから，本研究で設定した閾値 0.1 の時の適合率と再現率に注目すると，適合率は高いが，再現率は低くなってい

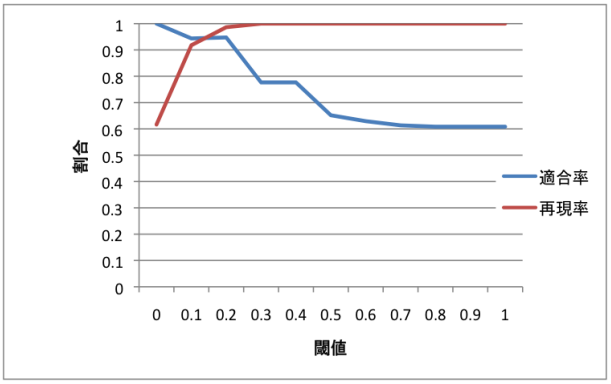

図 5: 閾値と適合率，及び再現率の関係のグラフ

る事がわかる．これは，有害文章と判別した文章はほぼ正しい判別ができているが，多くの有害文章を無害文章と誤判別している事を示す．有害情報のフィルタリングにおいて，有害文章を無害文章と誤判別するよりも，無害文章を有害文章と判別する方が危険度が低い．よって本来望ましい閾値は再現率が向上している 0.2〜0.3 の間で閾値を設定するのが好ましいと考えられる．このように，システムを使用する環境によって閾値による適合率や再現率の分析を行い，閾値を設定する必要がある．

4. 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，多くの有害情報の分類手法に関する研究がされている．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．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．また，井ノ上 [2] らは，URL チェックに加えてコンテンツチェック方式を組み合わせる手法有害情報のフィルタリングソフトの開発を行った．このコンテンツチェック方式のアルゴリズムは，パターン認識手法を用いて，不適切とはいえない表現も含めた単語・語句だけを登録する従来の手法とは異なり，VSM に基づいたベクトル各要素を IF*IDF として特徴ベクトルを抽出し，不適切とはいえない表現も含めた単語・語句の出現分布から Harzardous(有害情報) あるいは Safe(無害情報) のどちらかのカテゴリーに分類する文章自動分類手法を提案している．小林ら [3] は，知識検索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を行っている．ここでは，人手でフィルタリングされた投稿から，フィルタリングする際に暗黙的に用いられる知識を表出化し，フィルタリングの自動化と分類知識の共有を試みている．この研究では，単語間 (名詞のみ) の共起頻度が低いという性質を持つことから，文章をグラフ化し，Wikipedia を元に取得した正しい文書の共起頻度を比較することによって，その重なり具合が低い場合を禁止行為とし，フィルタリングを行っている．本稿の提案手法では，単語単体の出現確率ではなく，単語間の共起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実現する．

5. まとめと今後の課題

5.1 まとめ

本稿では 2 単語間の共起情報を元に，辞書データベースの構築を行い，辞書を利用した文章分類手法を提案した．また，実在する SNS の文章を用いた 2 つの評価実験を行った．性能実験では，9 割以上の精度で有害文章，及び無害文章の判別が可能である事を示した．比較実験では，既存の手法よりも，提案手法の方が文章の特徴をより抽出できている事を示した．また，本システムの精度は閾値の設定に依存する事から，閾値の変化による精度や適合率，及び再現率の分析を行い，適切な閾値を設定する必要がある事を示した．

5.2 今後の課題

以下に，本研究に関する今後の課題を述べる．

- 学習データの追加

  本研究で収集した学習データの数は約 15000 程であり，網羅的に判別を行うためには，さらなる学習データの収集が必要である．また，グレーワードやブラックワードに関しても，自動的に収集する方法を考慮する必要がある．具体的には，有害文章に含まれる単語の出現回数に対して閾値をもうけ，閾値以上，有害文章に出現した単語を抽出して，それらをグレーワード，もしくはブラックワードとして登録する手法等が考えられる．

- 特徴語の重み付け

  本研究においては，特徴語に関する重み付けを行っていない．有害文章と無害文章の共に出現する単語はあまり特徴を持っていないと考えられる．例えば「今日」や「私」などの単語は，有害文章にも無害文章にも出現する．そこで，より文章の特徴を抽出できる単語に関して TF-IDF を用いた特徴語の抽出，そしてその単語の重み付けを行う事で，さらなる判別精度の向上を目指す．

参考文献

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

[2] 井ノ上直己, 帆足啓一郎, 橋本和夫, ”文書自動分類手法を用いた有害情報フィルタリングソフトの開発 ”, 電子情報通信学会論文誌 D-IIVol. J84-D-II No. 6 pp. 1158-1166 June 2001.

[3] 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 ”, The 20th Annual Conference of the Japanese Society for Artificial Intellifence, 2006