Vol.2010-ICS-159 No.1
2010/3/15

２単語間の共起情報と距離情報を考慮した有害文章判別手法の提案

藤井　雄太郎[†1]　安藤　哲志[†1]　伊藤　孝行[†2]

近年，急激に発達している SNS 上には，未成年ユーザにとって有害な情報を配信するユーザが存在し，問題となっている．そこで本稿では，効率的に未成年ユーザに対して有害な文章をフィルタリングする事を目的とし，それらの文章を２単語間の共起情報を用いた判別方法により自動的に判別するシステムを提案する．また，実在している SNS の文章を用いて判定実験を行い，本システムの有効性を示す．

developing a system that filter harmful information for minors based on 2-word co-occurence information

Yutaro Fujii,[†1] Atsushi Ando[†1] and Talayuki Ito[†2]

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 minors into the services. This paper aims at developing a system that filter harmful information for minors 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，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．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．

†1 名古屋工業大学
Nagoya Institute of Technology
†2 MIT スローン経営大学院
MIT Sloan School of Management

2. 関連研究

2.1 有害情報フィルタリング

文書自動分類の中でも，ある特定の情報を抽出，排除するための技術としてフィルタリングがある．現在で，利用されている有害情報のフィルタリング方法は大きく２つに分類でき，１つは URL 方式，もう１つはコンテンツチェック方式である．本研究では，コンテンツチェック方式について関連する．

2.2 コンテンツチェック方式

コンテンツチェック方式では，利用者がアクセスしたコンテンツに含まれる単語・語句をチェックし，その結果に不適切な語句が含まれているコンテンツへのアクセスを制限する方式である．コンテンツチェック方式では，コンテンツに含まれる語句や単語をチェックして自動判定するため，立ち上がって間もないサイトを見つけて登録する等の作業が不要となる．しかし，精度が低い判定を行うと，無害なサイトや文章までもをフィルタリングしてしまう問題点があり，この方式の課題である．本研究における有害文章の判別システムの手法もコンテンツチェック方式に属する．以下に同様の関連研究を紹介すると共に，本研究の位置づけを述べる．

井ノ上[1)] 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．

Graham ら[?)] は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．こ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．

小林ら[2] 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，Yahoo!JAPAN の Yahoo!知恵袋を題材に，Yahoo!JAPAN がガイドラインに禁止行為として公表しているものの内,「Yahoo!JAPAN が予定していない目的で本サービスを利用している」投稿が用いられており，これらの投稿が単語間 (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章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．

本研究においては，有害文章の判別の対象として SNS，ブログ，及び掲示板等の文章としている．そのため，学習データとして，より対象の文章に文体が近しいブログ等から学習データを収集している．SNS 上に存在する文章はスラングや略語，話言葉が多く，文法的に成り立っていない文章が多い．一方，Wikipedia 等に存在する文章の文体は，日本語の文法的にも正しいものになっており，判別対象の学習データとして，適切であるとは限らない．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．

2.3 単語の定義

本稿で定義した単語を以下に説明する．
・ブラックワード bw:単体で有害な意味を持つ単語
・グレーワード gw:単語の使用の仕方によっては有害な意味も無害な意味も持ちうる単語．本稿では，共起の組み合わせ爆発を防ぐために用いる
・正例:学習データ用の例文．グレーワードを含み，その単語が無害な意味で用いら れている文章を指す．
・負例:学習データ用の例文．グレーワードを含み，その単語が有害な意味で用いられている文章を指す．

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[⋆1]，goo ブログ[⋆2]，2 ちゃんねる掲示板[⋆3]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．

⋆1 http://blogs.yahoo.co.jp/
⋆2 http://www.goo.ne.jp/
⋆3 http://www.2ch.net/

Vol.2010-ICS-159 No.1
2010/3/15

表 1 に辞書 DB の構造を示し，表 2 に辞書 DB のおおよそのデータ数を示す．

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	$l(gw,cw_i)\leq 5$ の出現回数 (無害文章)
dist_10_p	$6\leq l(gw,cw_i)\leq 10$ の出現回数 (無害文章)
dist_15_p	$11\leq l(gw,cw_i)\leq 15$ の出現回数 (無害文章)
dist_20_p	$16\leq l(gw,cw_i)\leq 20$ の出現回数 (無害文章)
dist_5_n	$l(gw,cw_i)\leq 5$ の出現回数 (有害文章)
dist_10_n	$6\leq l(gw,cw_i)\leq 10$ の出現回数 (有害文章)
dist_15_n	$11\leq l(gw,cw_i)\leq 15$ の出現回数 (有害文章)
dist_20_p	$16\leq l(gw,cw_i)\leq 20$ の出現回数 (有害文章)

表 2 辞書 DB のデータ数

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
共起の組み合わせ	8156156

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．

図 1 に，有害文章判別アルゴリズムの概念図を示す．

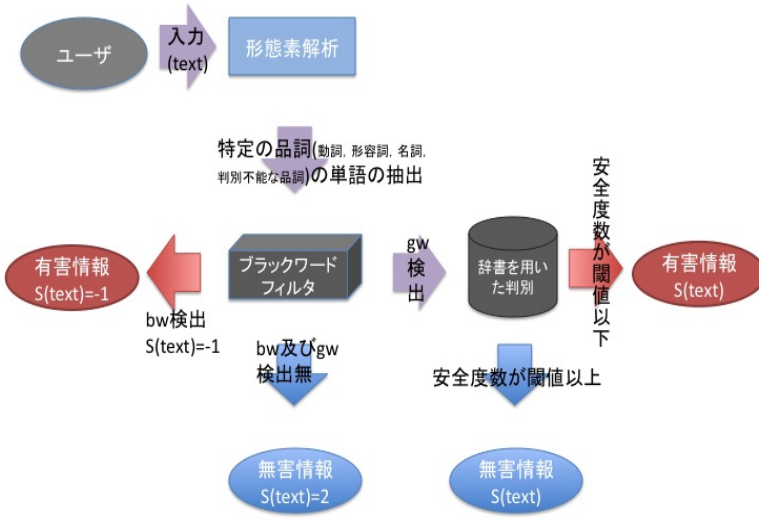

図 1 アルゴリズム概念図

$S(text)$ 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)〜(4) に計算式を示す．

・$l(gw,cw_i)$ ≤5 の時

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

・6≤ $l(gw,cw_i)$ ≤10 の時

$$s_i = \frac{dist\_5\_p + dist\_10\_p * 2 + \sum_{15} dist\_l\_p}{dist\_5\_p + dist\_10\_p * 2 + \sum_{15} dist\_l\_p + dist\_5\_p + dist\_10\_p * 2 + \sum_{15} dist\_l\_n} \quad (2)$$

・11≤ $l(gw,cw_i)$ ≤15 の時

$$s_i = \frac{\sum_{5}^{10} dist\_l\_p + dist\_15\_p * 2 + dist\_20\_p}{\sum_{5}^{10} dist\_l\_p + dist\_15\_p * 2 + dist\_20\_p + \sum_{5}^{10} dist\_l\_n + dist\_15\_n * 2 + dist\_20\_n} \quad (3)$$

・16≤$l(gw,cw_i)$ ≤20 の時

$$s_i = \frac{\sum_{5}^{15} dist\_l\_p + dist\_20\_p * 2}{\sum_{5}^{15} dist\_l\_p + dist\_20\_p * 2 + \sum_{5}^{15} dist\_l\_n + dist\_20\_n * 2} \quad (4)$$

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

そして，全ての単語 (cw_1,...cw_n) に対して s_i を計算する．最後に，式 (5) で，s_i の平均を計算し，その値を $S(text)$ とする．

$$S(text) = \frac{\sum_{i=1}^{n} s_i}{n} \quad (5)$$

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. 評価実験

4.1 実験方法

本実験では，試作したシステムにおける評価のため，評価実験を行った．実在する SNS 上の文章をテストデータとして，システムが適切に文章を判別できた精度として判別精度を求め，本システムを評価する．本実験では，SNS 上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．テストデータは yahoo 知恵袋[⋆1]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する．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，判別精度を明らかにする．無害な文章における判別精度 R_p((正判別した無害テストデータの数/無害テストデータ数)*100)，有害な文章における判別精度 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．

⋆1 http://chiebukuro.yahoo.co.jp/

Vol.2010-ICS-159 No.1
2010/3/15

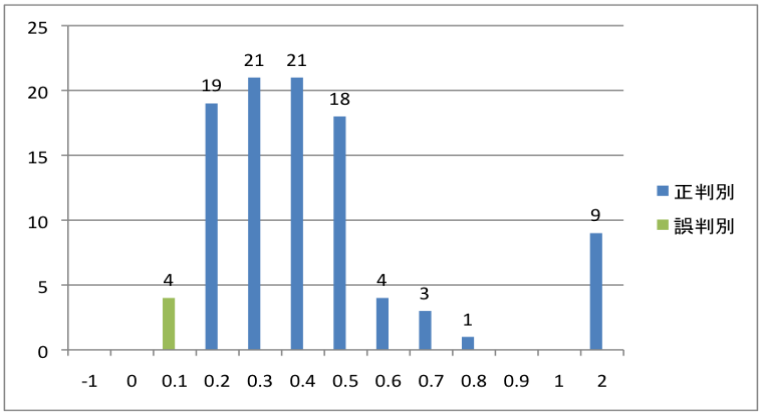

図 2 安全度数別の文章数の分布と判別結果

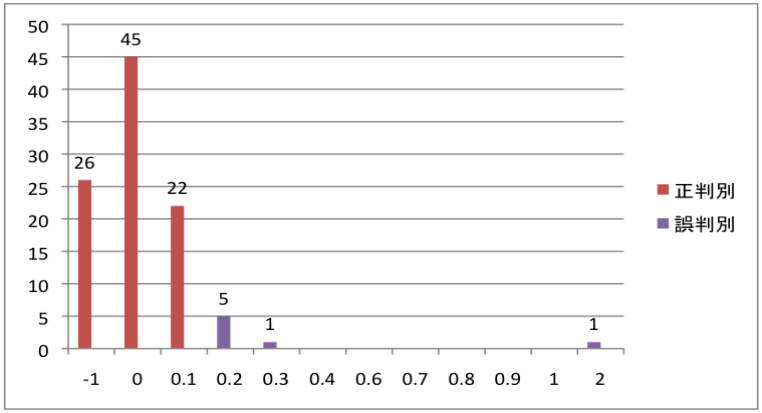

図 3 安全度数別の文章数の分布と判別結果

図 2 に実験結果として，無害テストデータの判別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判別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な文章，有害な文章ともに 90%以上の精度で判別する事がわかった．

4.2 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．

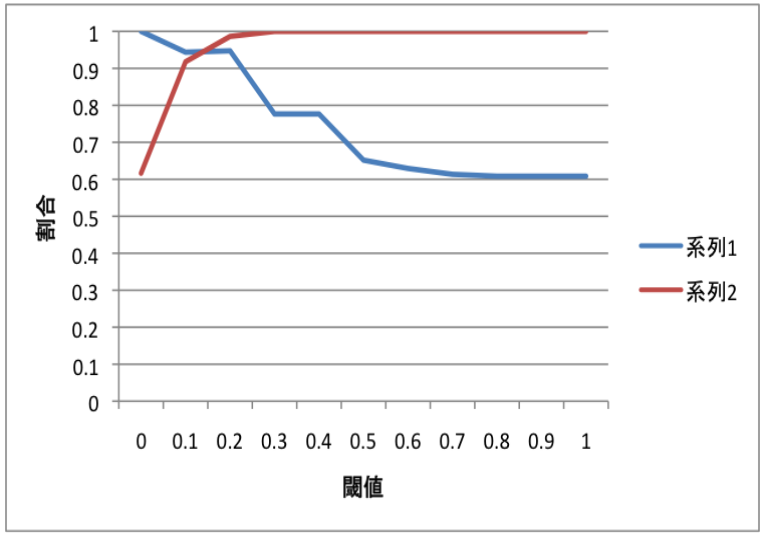

図 4 閾値と適合率，及び再現率の関係のグラフ

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．

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，辞書を利用した有害文章判別手法を提案した．また，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 割以上の精度で有害文章，及び無害文章の判別が可能である事がわかった．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．

- 形態素解析による単語分割方法
  本研究において，文章を形態素解析し，単語に分割する際に，名詞が連続で出現した時にはそれらの名詞を連結し，1 つの名詞として用いる方法をとっている．この方法では，連続して初めて意味が特定できる単語を抽出できるという長所がある．しかし，連続する事によって意味を失ってしまう単語もできてしまうという短所もある．つまり，ブラックワードのような単体で有害な意味を持つ単語と連続して名詞が出現し，それらが連結した単語が文法上の意味を成さないような単語ならば，そのような文章の判別が困難となる．この問題に対しては，全ての単語の区切りのパターンに対して，辞書を構築する手法が考えられる．

- 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．有害な文章と無害な文章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，有害な文章にも無害な文章にも出現する．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) 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.

2) 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

3) B. Grailheres, S. Brunessaux, P. Leray, Combining Classifiers for harmful document filtering, RIAO ’2004, Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval, 2004