3N-3

共起グループを用いた有害文書判定手法の提案

吉村卓也 † 藤井雄太郎 ‡ 伊藤孝行 †‡§
Takuya Yoshimura Yutaro Fujii Takayuki Ito

1. はじめに

Web によるコミュニケーションが急激に活性化しており，それにともない Web 上のコミュニケーションを提供するサービスを利用する人々も急増している．サービスは携帯電話からの利用が可能であるため，多くの未成年ユーザーも増えている．しかし，ユーザーの中には未成年にとって有害となる情報（暴力的な内容や，不適切な性的描写を含むものなど）を配信するものもおり問題となっている．有害コンテンツの対処が十分におこなわれておらず，多くは投稿されてから人手によって確認しており，管理側にかかる負担が大きくなるという問題がある．

本研究では，それらの人手による作業を自動化した情報のフィルタリングをおこなうことを目的とし，また，より高精度のフィルタリング手法の実装を目指す．実装するフィルタリング手法は，メールのスパムフィルタとしてよく利用されているベイジアンフィルタを基盤とする．提案手法では共起のグループ化をおこなう手法について述べ，純粋な共起を用いたフィルタリング手法より安定したフィルタリングを提案する．

評価実験ではフィルタリングの二つの方式である Paul Graham 方式と Gray Robinson 方式をもとに共起を用いたフィルタリング手法と共起グループを用いたフィルタリング手法の比較を行う．両方式で共起と共起グループでのフィルタリングの性能の評価し，フィルタリングの精度を示す．

2. 関連研究

Web 上には利用者側にとって害を及ぼす恐れのあるサイトも多く存在している．これらの有害サイトに制限をかけて情報のフィルタリングがおこなわれている．現在提供されているサービスでは現代的なフィルタリング手法を用いてられている．フィルタリングの対象となる要素をリストで管理する方式がある．アクセス制限をかける URL をリスト化したブラックリスト方式とリスト内に含まれる URL のみアクセスを許可するホワイトリスト方式である．また，一部の情報を隠蔽するストップワード方式があり，隠蔽される部分は文書の単語単位でおこなわれ，それらの単語をブラックワードよばれる．

ベイジアンフィルタは，ベイズ理論に基づいたフィルタリング手法で，判定後には対象に含まれる単語を新たに学習し，データを更新する特徴を持っている．この特徴から，それ以降に判定する対象に対する精度が向上していく．本研究では，Paul Graham 方式と Gray Robinson 方式の手法のフィルタリングを実装している．

Paul Graham 方式 [1] は，文書中に含まれる単語のなかから特徴的な単語を 15 単語を文書から抽出して，それら 15 単語をもちいて文書の有害確率を計算する．一度も出現したことのない単語に対しては一定の値を割り当てる．

Gray Robinson 方式 [2] では，ある一定の範囲内の有害率をもつ単語を除外して，それ以外の単語を文書の有害確率を計算するために用いる．また，Gray Robinson 方式は Paul Graham 方式ではうまく表現できなかった問題を事前確率を用いて表現可能としている．

また，ベイジアンフィルタを応用した共起フィルタリング [3] という手法がある．先に述べたベイジアンフィルタでは，単語の出現頻度をもちいて文書の有害度を計算するが，共起フィルタリングでは共起の出現頻度をもちいておこなう．

3. 共起グループを用いた有害文章判定手法

先に述べた既存手法のうち，Gray Robinson 方式では，計算に用いる要素を範囲の設定で制限するため，文書内のすべての単語が計算に含まれない可能性が生じるという問題がある．本問題を，本論文では，共起グループのすべての特徴量をもちいて判定を行うことで解決する．また，共起をグループ化することで，ある単語の文書内に含まれる他単語との関連性の特徴を抽出することができ，より複雑な判定が可能となる．共起グループとはある単語を含む共起の集合のことを表す．この共起グループを用いて文書の有害確率を計算する．以下に共起グループの有害確率の式を示す．

$$p(G_i) = \frac{1}{n}\sum_{i=1}^{n} p(c_i) \tag{1}$$

式 1 の $p(G_i)$ を用いて Paul Graham 方式と Gray Robinson 方式で文書の有害判定をもとめる．提案手法では，得られる結果はすべての共起の特徴量を考慮していることから，通常の共起フィルタリングより分散するため，曖昧な文書の判定が可能となる．

式 2 に共起グループを用いた Paul Graham 方式によるフィルタリングでの文書の有害確率を計算する式を示す．

$$p(D) = \frac{\sum_{i=0}^{n} p(G_i)}{\sum_{i=0}^{n} p(G_i) + \sum_{i=0}^{n}\{1 - p(G_i)\}} \tag{2}$$

以下 Gray Robinson 方式による共起グループを用いたフィルタリングの式である．

†名古屋工業大学情報工学科 Department of Computer Science, Nagoya Institute of Technology
‡名古屋工業大学大学院産業戦略専攻 Master of Technology Business Administration, Nagoya Institute of Technology
§東京大学政策ビジョン研究センター Policy Alternative Research Institute, University of Tokyo

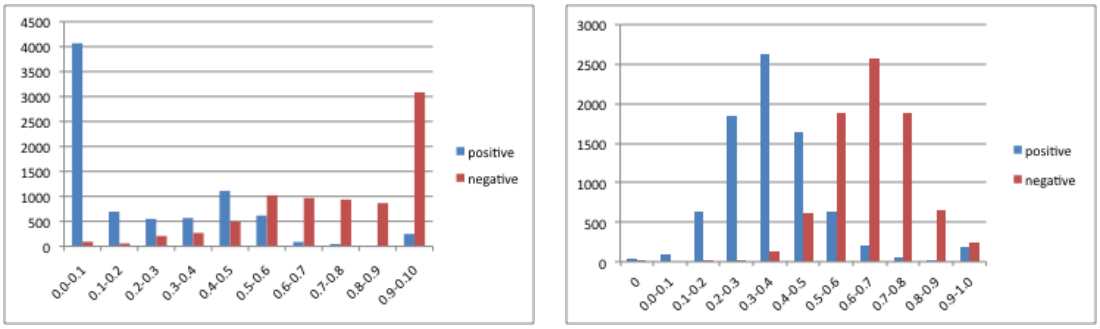

図 1.1: 共起フィルタリングと共起グループフィルタリング

表 1.1: フィルタリング効率の評価（F 値）

	無害文書に対する F 値	有害文書に対する F 値
Paul Graham 方式＋共起	0.859	0.853
Paul Graham 方式＋共起グループ	0.864	0.858
Gray Robinson 方式＋共起	0.868	0.866
Gray Robinson 方式＋共起グループ	0.878	0.884

$$f(G_i) = \frac{x \cdot s + n(G_i) \cdot p(G_i)}{s + n(G_i)} \tag{3}$$

$$n(G_i) = \frac{1}{l} \sum_{k=0}^{l} n_k \tag{4}$$

$$H(D) = 1 - \prod_{i=0}^{n} p(G_i)^{\frac{1}{n}} \tag{5}$$

$$S(D) = 1 - \prod_{i=0}^{n} \{1 - p(G_i)\}^{\frac{1}{n}} \tag{6}$$

式 3 は各共起グループの有害率を示し，式 5，6 は文書のスパム性のノンスパム性を表す．これら二つの値を用いて指標 I_1 をもとめて判定を行う．また，式 4 は共起グループ内に含まれるすべての共起の出現頻度の平均をあらわす．

$$I_1 = \frac{S(D) + H(D)}{S(D) - H(D)} \tag{7}$$

4. 評価実験

評価実験は，共起と共起グループを用いたフィルタリングの二つの手法の比較をおこない性能を F 値を用いて評価する．各フィルタリングの学習データは有害文書と無害文書を各々1 万件ずつ，サイズにすると 66.3MB のデータをもちいる．テストデータは，無害文書と有害文書を各々8000 件で，無害文書に対する再現率と適合率，有害文書に対する再現率と適合率を求め，再現率と適合率の調和平均 F 値で性能評価を行う．また，学習データとテストデータは 2 ちゃんねるから収集した．表 1.1 に実装したフィルタリングによる結果を示す．表 1.1 からわかるように，無害文書に対する F 値がもっとも高いのが Gray Robinson 方式と共起グループの組み合わせのフィルタリングとなる．同様に，有害文書に対する F 値も Gray Robinson 方式と共起グループの組み合わせが最も高い値をとなる．また，最低値はともに Paul Graham 方式と共起の組み合わせとなる．また，図 1.1 では共起グループと共起のフィルタリングの判定結果をグラフ化したものである．共起は左右に分かれ，共起グループは無害は 0.3 から 0.4，無害は 0.6 から 0.7 の間の分布を頂点とした山の分布となっている．

5. まとめ

人手でおこなわれるフィルタリングはコストがかかる問題がある．その問題を解決するために，自動的にフィルタリングすることが重要視されている．本研究では，ベイジアンフィルタに基づいたフィルタリングを実装することで人手の作業コスト削減を実現した．また，より精度の高いフィルタリングを実装するため，既存手法の共起フィルタリングと提案手法である共起グループフィルタリングの比較実験を行った．実験結果から，共起グループによるフィルタリングは既存手法より高精度の性能があることがわかった．

参考文献

[1] Paul Graham. A plan for spam, In P. Graham, Hackers and Painters. O'Reilly. O'Reilly, 2004.

[2] Robinson Gray. Spam detection, 2002. http://radio.weblogs.com/0101454/stories/2002/09/16/spamDetection.html.

[3] 安藤哲志，藤井雄太郎，伊藤孝行．“有害文書判別のための他単語間共起情報辞書の構築とその応用”，2010．情報処理学会第 72 回全国大会．