RF-007

Robinson 型判定手法を用いた単語共起フィルタの検証
Verifying Co-occurrence Filtering Based on Robinson Type Filter

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

1. はじめに

近年，掲示板や Social Network Service(SNS) のようなユーザが自由に読み書きのできるコミュニケーションツールが Web サイト上に増えてきている．しかし，Web 上に存在する様々な情報の中には悪意のある情報，例えば未成年にとって悪影響を及ぼす書き込みなどが存在し，問題となっている．2006 年に携帯電話事業者に対して総務省から有害サイトアクセス制限サービスの普及促進の要求 [1] がおこなわれており，有害文書の対処が必要とされてきている．しかし，多くの Web サイトでは有害情報に対する対策をおこなわれておらず，また対策をしている Web サイトにおいても，情報が発信された後に人の目視による確認で対処がなされている．発信された後に人が直接確認する対処方法では時間も手間もかかるため，情報が膨大になるにしたがって処理が追いつかない問題が生じている．そのために，現在は自動的に有害な情報をフィルタリングするための研究が多くされている．

従来のフィルタリング手法として，ブラックリスト方式，ホワイトリスト方式，ストップワード方式などが挙げられる．ブラックリスト方式では，アクセス制限の対象となる有害な内容を含んだ Web サイトの URL をリスト化する方法である．ホワイトリスト方式では，リスト化された URL のみのアクセスを許可する方法である．ストップワード方式は、ブラックワードと称される有害な意味をなす単語をリストにし、ブラックワードを含む Web サイトにアクセス制限をかける方法である．

ブラックリスト方式やホワイトリスト方式によるフィルタリングでは SNS 全体にアクセス制限がかかる恐れがある．すべての URL を人手で行わなければならない事から，ユーザの用いる単語の変化が著しい現在の Web 上では対応しきれない．また，ストップワード方式ではブラックワードの選定基準等に大幅なコストがかかるため効率的ではない．スラングや隠語などが含まれた文書に対しては適切な対応ができないため，昨今の Web 上に今述べた三つのフィルタリングの手法で対応し続けるのは困難である．

本稿では，文書の特徴を抽出し，有害か無害かを自動判別する手法についていくつか紹介し，既存のフィルタリング手法を応用した，共起のグループ化によるフィルタリング手法を提案し比較実験をおこなう．三つの既存のフィルタリング手法と比較実験をおこない，提案した手法の有用性について考察する．

実装したフィルタリングは，統計的学習手法に基づいて [2]，スパムメールのフィルタリングに用いられている Paul Graham[3, 4] や Gray Robinson[5, 6] が考案したベイジアンフィルタ [7, 8, 9, 10, 11, 12] による手法を基盤としている．Paul Graham 方式によるスパムメールフィルタリングでは，単語毎のスパム確率を求めてより特徴的な単語のうち 15 単語を用いて全体のスパム確率を計算する．単語毎のスパム確率を求める際に，バイアスをかけて無害文書の誤検出率を低くする工夫や，一度も出現した事のない単語や，極端に出現頻度の少ない単語に対しては一定の確率を与える工夫が施されている．一方，Robinson によるフィルタリングでは，同様に単語毎のスパム確率を求めるが，単語の出現頻度の公平性を保つために，出現頻度関数の解を求めて判定基準の指標を計算する．また，Graham 方式によるフィルタリングより Robinson 方式によるフィルタリングの方が誤検出率が低い結果が発表されている [11]．

本稿では，第２章にて関連研究について述べて，第３章で学習データの構築について、第４章で共起フィルタリングの実装について説明し，第５章でまとめる．

2. 関連研究

2.1. 従来のフィルタリング

有害サイトのフィルタリングサービスは検索ポータルサイトなどから既に多く提供されている．Yahoo!が提供している Yahoo!あんしんねっと [13] では，ブラックリスト方式とホワイトリスト方式に加えキーワードによるフィルタリングが採用されており，利用者側でこの三つの方式を選択できる仕組みになっている．キーワードフィルタリングでは，サイト内に現れる単語で不適切な単語を伏せ字に置き換えて，有害な情報を閲覧できないようにするものである．しかし，Yahoo!が提供しているサービスには先に挙げた、アクセス制限の境界や人手によるコストの問題点が依然として残っている．

2.2. ベイジアンフィルタ

ベイジアンフィルタはスパムメールのフィルタリングに有効な手段として用いられていて，単純なベイズ分類器を応用して，対象のデータを解析して学習，分類をおこなうためのものである．応用例として Paul Graham 方式や Gray Robinson 方式などがあげられ，本稿では，今挙げた二種類の手法を基盤にフィルタリングを実装している．

1. Paul Graham 方式

†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工学専攻 Master of Techno-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

Paul Graham 方式では対象文書内の各単語のスパム確率を求めていく．無害文書の誤検出率を下げるため，正例の学習データから得られる出現回数にバイアスをかける．文書のスパム確率は次の式で求める．

$$p(w_i) = \frac{\frac{b_i}{N_{bad}}}{a \times \frac{g_i}{N_{good}} + \frac{b_i}{N_{bad}}} \quad (1)$$

$$p(D) = \frac{\prod_n^{i=1} p(w)}{\prod_n^{i=1} p(w) + \prod_n^{i=1} 1 - p(w)} \quad (2)$$

$P(w_i)$ は単語 w_i のスパム確率を表す．b_i は負例の学習データから得られる単語 w_i の出現回数を示し，g_i は正例の学習データから得られる単語 w_i の出現回数を示しており，N_{good} と N_{bad} はそれぞれ正例負例の学習データから得られる総出現回数を表している．単語 w_i が学習データに存在しない単語，あるいわ，$2gi + bi \leqq 5$ の条件を満たす出現回数が極端に少ない単語の場合には確率 0.4 を与える．各単語の (1) 式で得られる値から特徴的な単語を 15 単語を抽出する．抽出する基準は，スパム確率が 0.5 との絶対値の差の降順とする．抽出した特徴語 15 単語を用いて次の式の値を指標にスパムの判定をおこなう．また，g_i にはバイアス値 a がかけられており，無害文書の誤検出率を下げている．$P(D)$ が 0.9 以上のとき対象の文書を有害と判定し，それ以外は無害と判定する．

2．Gray Robinson 方式

Paul Graham 方式では過去に出現した事のない単語や，極端に少ない出現回数の単語に対して一定の確率を与えていたが，総出現回数に応じてスパム確率に重みを与えられるように改善された確率モデルを使ったフィルタリングである．単語毎に与えるスパム確率は以下のようになる．

$$p(w_i) = \frac{\frac{b_i}{N_{bad}}}{\frac{g_i}{N_{good}} + \frac{b_i}{N_{bad}}} \quad (3)$$

$$f(w_i) = \frac{s \cdot x + n \cdot p(w_i)}{n + s} \quad (4)$$

確率 $p(w_i)$ は Graham 方式の (1) 式のバイアスをかけない式に等しい．Robinson 方式ではバイアスをかけない (3) 式から新たに出現頻度関数を定義する (4)．s は過去一度も出現した事のない単語に対する強さを表し，n は単語 w_i の総出現回数 (正例で出現した回数と負例で出現した回数の和) を示す．x は事前確率を表し，本稿では，事前確率 x は判定対象の文書内にあるすべての単語のスパム確率の平均値である．$f(w_i)$ は単語の出現回数 n が極端に少なく，かつ $p(w_i)$ が高いときにスパム確率を下げることができるため，無害文書の誤検出率を下げることができる．(4) 式を使い，(5:eql), (6) 式よりスパム性 $S(D)$ とノンスパム性 $H(D)$ を求めて，$S(D)$, $H(D)$ を用いて判定指標 I_1 を求める．また，指標を 0 から 1 の範囲で判定するために指標 I_1 を指標 I_2 に変換して，指標 I_2 を用いて文書判定をおこなう．

$$S(D) = 1 - \prod_{i=1}^{n} \{(1 - f(w_i)\}^{\frac{1}{n}} \quad (5)$$

$$H(D) = 1 - \prod_{i=1}^{n} f(w_i)^{\frac{1}{n}} \quad (6)$$

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

$$I_2 = \frac{1 + I_1}{2} \quad (8)$$

2.3. 共起フィルタリング

共起フィルタリングは文章中に含まれる複数の単語が共起する組み合せを考慮するフィルタリングである．ベイジアンフィルタでは学習データから得られる単語の出現頻度を用いてフィルタリングをおこなうが，共起フィルタリングでは共起の出現頻度を用いておこなう．本稿では，二単語共起によるフィルタリングを実装している．

3. データベース

本稿では Web 上から収集した正例と負例の学習データから形態素解析ツールを用いて共起情報を抽出して共起辞書を構築する．学習データの内容や収集手段，共起辞書構築までのシステム構成について解説する．

3.1. データモデル

計算コスト削減のために単語および共起の ID ハッシュ化をおこなう．分割された単語から ID 化しデータベースを構築し，その ID 化された単語のデータベースを用いて，共起関係の ID 化をおこなう．

(例)
- ・文書　[今日は，良い天気だ．]
- ・単語分割　[今日, 良い, 天気]
- ・単語の ID 化　[単語] → [ID]
  [今日] → [1]
  [良い] → [34]
  [天気] → [102]
- ・共起関係の ID 化　[共起] → [共起 ID]
  [1, 34, 102] → [3]

表 1: 単語インデックス

データ属性	データタイプ
単語 ID	STRING
ID	INT

表 2: 共起インデックス

データ属性	データタイプ
共起 ID	INT
単語 1	INT
単語 2	INT

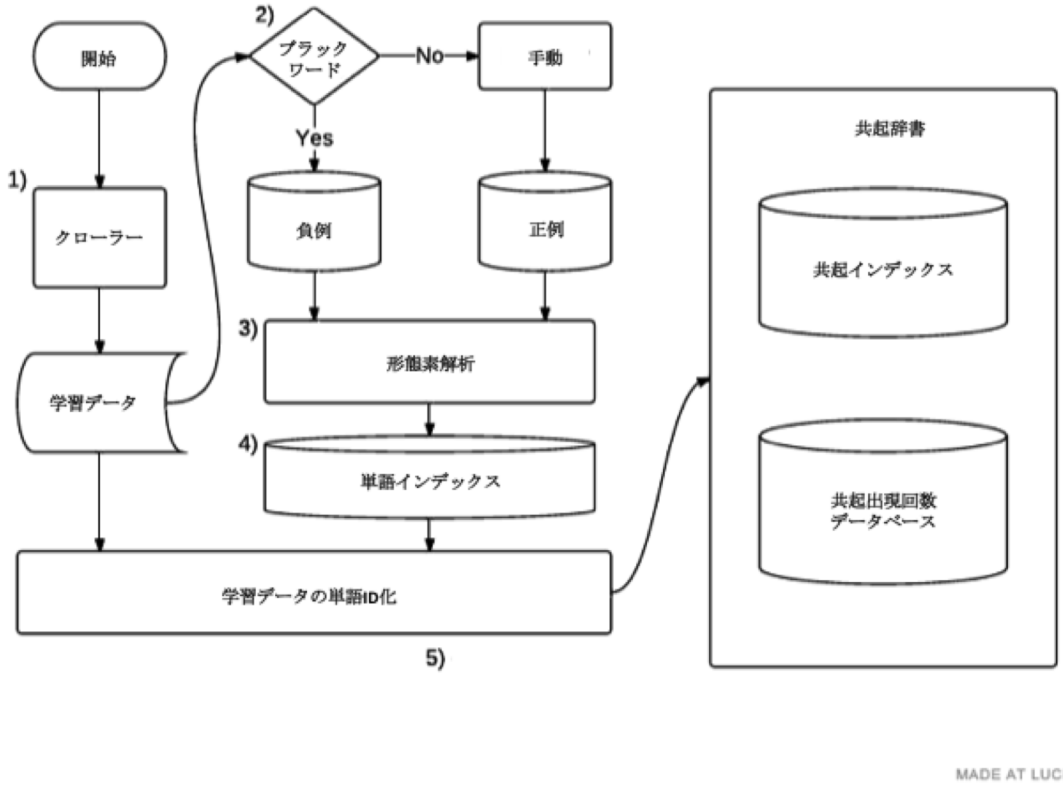

図 1: 共起辞書構築フローチャート

表 3: 共起出現頻度データベース

データ属性	データタイプ
共起 ID	INT
正例出現回数	INT
負例出現回数	INT

3.2. 学習データの収集

学習データの収集は，Web 上の Yahoo!ブログ，2 ちゃんねる等の掲示板の文書をクローラーを用いて行った．本稿で，1 件のデータサイズが 3.5KB で，合計 66.3MB の 2 万件の学習データを扱っている．収集した学習データを単語へ分割をおこなうために形態素解析ツール MeCab[14] を用いて，文書を単語へと分割していく．分割する際には助詞などの単体で意味を成さない単語は抽出をしない．抽出する単語数を削除することで計算コストを削減することが主な目的であり，単体で意味をなさない単語は取り除いても精度に影響は出ないとされている [7].

3.3. 共起辞書構築までの流れ

図 1 は二単語共起フィルタリングを行うためのデータベース構築までのフローチャート図である．1）で Web クローラーを用いて Web 上にある文書を収集し，学習データとして扱う．2）ではストップワード方式による文書のフィルタリングをおこない，収集したデータを有害と無害に分割する．3）は正例と負例に分類した文書を形態素解析して単語に分割し，4）で単語インデックスを生成する．単語インデックスと学習データを用いて ID 単語列に変換したテキストファイルを生成して，5）で共起を ID にハッシュ化した共起インデックスと共起の出現頻度のデータベースからなら共起辞書を構築する．ID 単語列に変換するのは共起インデックスを単語 ID で参照できるようにするためで，INT で

4. 提案手法の実装

文書を有害か無害に判定するために共起を用いた Gray Robinson 方式によるフィルタリングをおこなうと，共起の組み合せ数が多すぎて計算に誤差が生じて，判定が困難となる．共起の組み合せ数による計算の誤差をなくすため，同じ単語を含む共起をグループ化し，単体の共起を用いるのではなく，共起のグループを用いて計算コストを削減するフィルタリングを実装した．共起に着目したとき，Paul Graham 方式では特徴語 15 単語の spam 確率から指標を計算するが，Gray Robinson 方式ではすべての単語の spam 確率を考慮した式になるため，共起の組み合せ数が多いと (5) 式と (6) 式の積和の部分が情報落ちにより 0 になり，$H(D)$ と $S(D)$

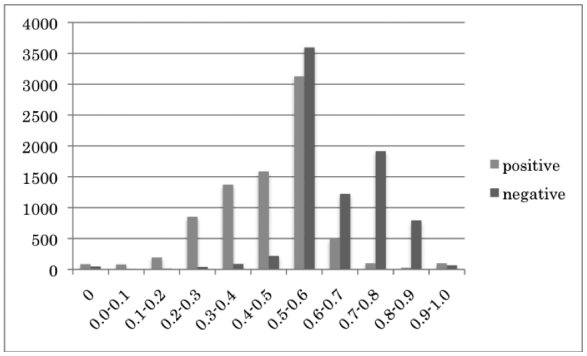

図 2: 0.5 に収束するグラフ

の値が 1 になる恐れがある．(8) 式で指標 I_2 を求めると値が 0.5 に収束してしまい，フィルタリングとしての機能が働かなくなる．情報落ちの誤差による機能の欠陥を改善するために，共起の組み合せ数分の積和を求めるのではなく，特定の同じ単語を含む共起の集合 (共起グループ) の特徴量の積和を求めることで計算コストを削減する．共起グループの特徴量は，(9) 式で求める．

$$f_A = \frac{s.x + n_{Aave}}{n_{Aave} + n} \tag{9}$$

$$n_{Aave} = \sum_{i=1}^{n}(g_i + b_i) \tag{10}$$

$$P_{Aave} = \sum_{i=1}^{n} p(w) \tag{11}$$

f_A はグループ A の出現頻度関数を表し，f_A は (4) 式の単語 w_i の出現頻度 n と spam 確率 $p(w_i)$ の値を共起グループ内の共起出現頻度の平均 n_{Aave}s と pam 確率の平均 p_{Aave} に置き換えた式になる．平均値をとることで，文書の形態素数で積和の計算コストを抑えることができる．

5. 実験結果

Paul Graham 方式によるベイジアンフィルタと共起フィルタ，Gray Robinson 方式によるベイジアンフィルタと共起フィルタについて比較実験をおこない，各フィルタリングの特徴と違いについて考察する． Graham 方式によるフィルタリングでは，閾値は 0.9 で (2) 式で得られる文書の特徴量が閾値より大きいとき有害と判定する．図 5 から，無害文書は 0 から 0.1 の範囲の値に，有害文書は 1 に近い値に集中していることがわかる．図 6 共起フィルタも同様に，無害文書と有害文書の判定は一極端に集中している．無害文書に着目すると，共起フィルタの方が 0 から 0.1 の間に分布する値が多く，Graham 方式の特徴が顕著にあらわれている．3 と 4 を比較しても，無害文書のフィルタリングの判定率が高いことがわかる．しかし，有害文書の判定率は低くなるため，表 8 の調和平均 F 値がベイジアンフィルタの方が高い値を出している．結果としては，精度はベイジアンフィルタの方が高いことがわかるが，図を比較すると共起フィルタの方が Graham 方式の特徴が顕著に出ている．精度が悪い原因として，必要な学習量の違いが挙げられる．ベイジアンフィルタでは単語のみを学習するが，共起フィルタでは共起の組み合せのを学習するため，組み合せの分だけ学習量が必要になる．

表 4: Paul Graham 方式によるベイジアンフィルタ

	無害判定数	有害判定数
無害文書	6968	1032
有害文書	1190	6810

表 5: Paul Graham 方式による二単語共起フィルタ

	無害判定数	有害判定数
無害文書	7138	862
有害文書	1420	6580

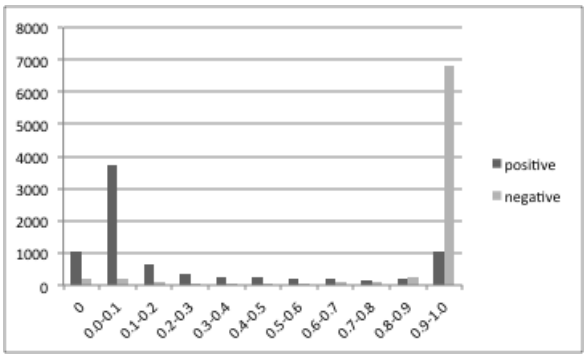

図 3: Paul Graham 方式によるベイジアンフィルタ

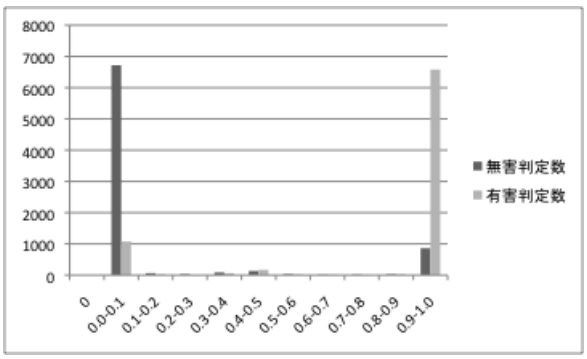

図 4: Paul Graham 方式による共起フィルタリング

Robinson 方式によるフィルタリングの場合，閾値は 0.5 で (8) 式で求める指標 I_2 の値が閾値より大きければ有害として判定する．図 7 のベイジアンの結果では有害文書は 0.6 から 0.7 の範囲の値が最も多く，無害文書は 0.4 から 0.5 の範囲の値が最も多い．共起フィルタでもフィルタリングをおこなった結果，図 8 に示されるように図??のときよりも分布が分散していることがわかる．また，閾値周辺の値の分布数が増加しているため，偏りのある判定結果となる，表 5 から判定数を

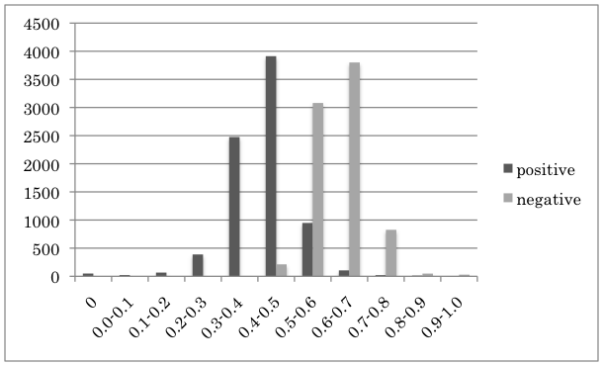

図 5: Gray Robinson 式によるベイジアンフィルタ

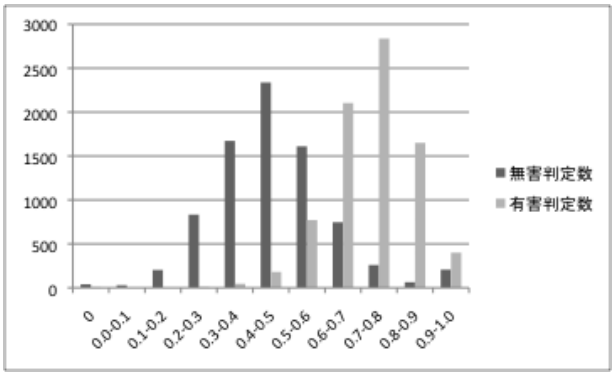

図 6: バイアス a=1 のとき

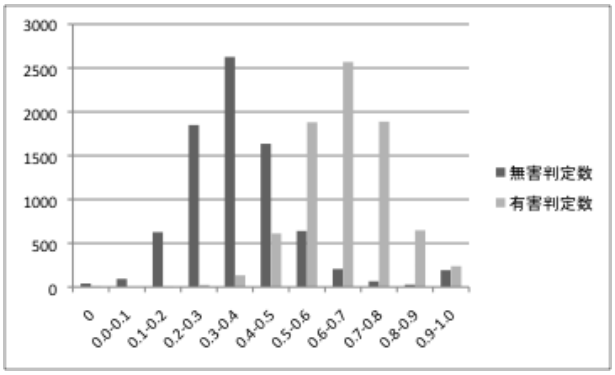

図 7: バイアス a=2 のとき

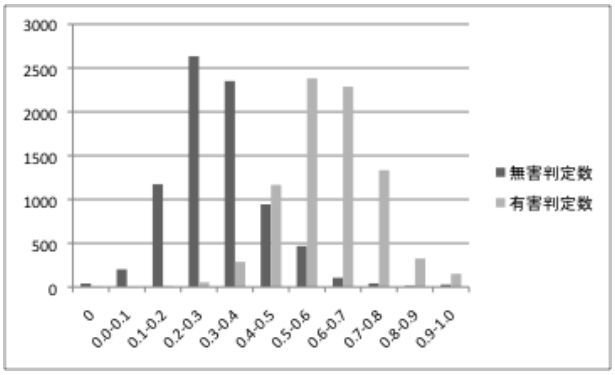

図 8: バイアス a=3 のとき

みると，有害文書に対する判定率が高くなり，無害文書の誤検出率も増加している．無害文書の誤検出率を下げるため，Graham 方式と同様にバイアス値をかけた (1) 式を用いて実験をおこなう．バイアス値 $a = 2$ のとき，図 9 の結果が得られる．閾値の周辺がグラフの谷になる分布を示し，表 6 から無害判定の誤検出率が下がり，精度が向上したことがわかる．また，バイアス値 $a = 3$ で実験をおこない比較したところ，図 10 をみると，図 10 と図 10 の分布と異なり，グラフの山の頂点が右に幅が広がるため，表??の判定数の結果をみると，バイアス値 $a = 2$ のときよりも無害判定の誤検出率が高くなる．Robinson 方式ではバイアスをかけないが，共起グループによるフィルタリング手法を用いたことで，副作用が生じた可能性がある．

実装したすべてのフィルタリング手法の結果をみると，表 8 から，Robinson 方式によるベイジアンフィルタと共起フィルタ ($a = 2$) の精度が高いことがわかる．しかし，再現率と適合率の格差をみると共起フィルタの方が少なく調和がとれている．精度は同程度だが，共起フィルタの方が安定したフィルタリングができることがわかる．また，ベイジアンフィルタと共起フィルタの比較をしたとき，共起を用いれば必ずしも精度が向上する訳ではないが，グラフでの値の分布は Graham 方式と Robinson 方式の特徴が顕著にあらわれる傾向がある．精度が向上しない原因として，先にも挙げた必要とする学習量の違いである．

6. まとめ

本稿では，単語共起のデータベースの構築方法と，Gray Robinson 方式によるベイジアンフィルタを基盤とした，共起のグループを用いたフィルタリング手法の実装について解説した．特定の単語を含む共起の集合の特徴量を抽出して計算コストを削減することで誤差をなくす手法である．既存のフィルタリング手法として，Paul Graham 方式によるベイジアンと共起のフィルタリング，Robinson によるベイジアンフィルタを実装し，本稿で提案した手法との比較実験をおこなった．結果として，Gray Robinson 方式によるベイジアンフィルタと共起フィルタの精度が高く，共起フィルタの方が安定したフィルタリングがおこなえることがわかった．Robinson 方式を用いた場合，分布の値が分散するため，多段フィルタに有効な手法としても用いることができる．また，Robinson による共起フィルタではバイアスの値を調整しているが，本来バイアスをかけないため，共起グループによる手法で副作用が生じた可能性があるため，検証が必要である．比較実験の結果，共起をとることで必ずしも精度が上がるという結果とならず，一つの原因として学習量の違いによるものが挙げられる．

表 6: Gray Robinson 方式によるベイジアンフィルタ

	無害判定数	有害判定数
無害文書	6415	1585
有害文書	495	7505

表 7: バイアス a=1 のとき

	無害判定数	有害判定数
無害文書	5114	2886
有害文書	239	7761

表 8: バイアス a=2

	無害判定数	有害判定数
無害文書	6870	1130
有害文書	778	7222

表 9: バイアス a=3 のとき

	無害判定数	有害判定数
無害文書	6415	1585
有害文書	495	7505

表 10: 再現率と適合率と F 値

	再現率	適合率	F 値
ベイジアンフィルタ (Graham)	86.8%	85.1%	0.859
ベイジアンフィルタ (Robinson)	82.6%	95.1%	0.884
二単語共起+Graham	88.4%	82.2%	0.852
二単語共起+Robinson (a = 1)	72.9%	97.0%	0.832
二単語共起+Robinson (a = 2)	86.4%	90.3%	0.883
二単語共起+Robinson (a = 3)	82.6%	93.8%	0.878

参考文献

[1] 青少年が使用する携帯電話・phs における有害サイトアクセス制限サービス（フィルタリングサービス）の導入促進に関する携帯電話事業者への要請. http://www.soumu.go.jp/menu_news/snews/2007/071210_4.html.

[2] Schutze H. Foundations of statistical natural lanperspectives. New York: Oxford Univ. Press, 1999. guage processing. Cambridge, MA: MIT Press, 1999, 1999.

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

[4] Paul Graham. Better bayesian filtering. proceedings of the 2003 spam conference, 2003.

[5] Robinson Gray. A statistical approach to the spam problem. http://www.linuxjournal.com/article/6467/.

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

[7] 安藤哲志, 藤井雄太郎, 伊藤孝行. “有害文書判別のための多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.

[8] 津田裕一, 八木秀樹, 平澤茂一. “単語の共起を考慮に入れたナイーブベイズモデルによる文書分類”, 2006. 第 29 回情報理論とその応用シンポジウム予稿集.

[9] 谷岡広樹, 中川尚, 丸山稔. “迷惑メールフィルタのためのベイジアンフィルタの改良”, 2007.

[10] 菊池琢弥, 内海彰. “語の共起情報に基づく有害サイトフィルタリング手法”, 2010. 第 9 回情報科学技術フォーラム (FIT2010) 講演論文集.

[11] 王卉歓, 中谷直司, 小池竜一, 厚井裕司. “ベイズアルゴリズムのスパムフィルタとウィルスフィルタへの適用の最適化 (侵入検出・見地,¡特集¿情報システムを支えるコンピュータセキュリティ技術の再考)”.

[12] 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, Vol. 1.J84-D2, No. 6, 2001.

[13] Yahoo!あんしんネット. http://anshin.yahoo.co.jp/.

[14] Mecab. http://mecab.sourceforge.net/.