RD-001

語の共起情報に基づく有害サイトフィルタリング手法

Harmful Website filtering based on word co-occurrence information

菊池 琢弥† 内海 彰 †
Takuya Kikuchi Akira Utsumi

1 はじめに

近年，インターネットは急速に普及し，その利便性から，我々の生活において必要不可欠なものとなっている．最近では携帯電話によるインターネットの利用も広まり，パソコンを持っていなくともインターネットを容易に利用できるようになっている．今日では，我が国でのインターネットの利用率は 75%であり，13〜19 歳に限れば 95%にのぼる [1]．このように，若年層がインターネットを利用する割合はかなり高くなっている．

しかし，インターネットから得られる情報には，青少年にとって有害な内容を含んでいるものも少なくない．特に最近では，子どもが「出会い系サイト」といった青少年にとって有害なサイトにアクセスすることで事件に巻き込まれるケースが多発している．

財団法人インターネット協会では，青少年の安全なインターネット利用に関するインターネット上のコンテンツの格付け基準として，SafetyOnline3[2] を策定している．これは，先に述べたように有害サイトによる被害が社会問題化していることから，有識者，関連事業者からなるレイティング／フィルタリング連絡協議会研究会によって検討され，取りまとめられたものである．その一部を表 1 に示す．

総務省では有害サイトのフィルタリングの普及促進を進めており，これを受け，各携帯電話事業者は，18 歳未満の携帯電話利用者へのフィルタリングサービスへの加入を原則義務化している．さらに，青少年の安全なインターネット利用を目的とし，携帯電話事業者にはフィルタリングサービスの提供，パソコンメーカーにフィルタリングソフトを組み込みを義務化する「有害サイト規制法」が成立するなど，青少年の安全なインターネット利用に向けた動きは活発化し，有害サイトのフィルタリングの重要性は高まっている．

フィルタリング方法としては，従来より，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使われており，これらに基づいたフィルタリングが主流である．ブラックリスト方式では，リストに登録された URL をフィルタリング対象とするもので，ホワイトリスト方式ではブラックリスト方式とは反対に，リストに登録された URL のみ閲覧可能とするものである．ストップワード方式は，禁止単語を設定し，その単語を含むサイトをフィルタリング対象とするものである．

しかし，ブラックリスト方式やホワイトリスト方式では，各リストの作成および管理は人手で行う必要があり，日々変化していくインターネット上のサイトに素早く対応することは不可能である．ストップワード方式においては，一度ストップワードリストを作成してしまえば，人手での確認作業が必要ない．そのために先に述べたような即応性の問題はないが，この方法においてもリストの作成は人手で行うため，どの語をストップワードとするのかが大きな問題となる．

以上の問題に対して，人手ではなく，有害サイトの特徴を検知し，それに基づいてフィルタリングを行う手法が求められている．本論文では，迷惑メールのフィルタリングに多く用いられているベイジアンフィルタ [5] を有害サイトのフィルタリングに応用し，さらに 2 語の共起による有害確率を考慮することによって，有害サイトを高精度に判定する手法を提案する．本論文の以下では，2 章で関連研究および既存のフィルタリングサービスについて，3 章では既存のベイジアンフィルタの手法について述べる．4 章では本論文での提案手法について，5 章では今回行った評価実験について述べ，6 章で結果について考察し，7 章でまとめる．

2 関連研究

2.1 既存フィルタリングサービス

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で既に多く提供されている．

検索サイト Goo では，子ども向けフィルタリングサービスとしてキッズ goo[1]を提供している．これは，ブラックリスト方式とストップワード方式を併用していると思われる．また，Yahoo が提供する Yahoo!あんしんネット [3] では，ブラックリスト方式，ホワイトリスト方式の 2 つに加えて

表 1: SafetyOnline3 の内容 (抜粋)

区分	年齢区分
ヌード	18 歳未満閲覧禁止
性行為	
性風俗情報	
暴力表現	
格闘（格闘技を除く）	
恐怖表現	
不快表現	
薬物・劇薬物	
差別的表現	
ギャンブル	
飲酒・喫煙	
出会い	
自殺	
参加型サイト	15 歳未満閲覧禁止
チャット	
ショッピング	

†電気通信大学大学院情報理工学研究科総合情報学専攻

[1]キッズ goo：http://kids.goo.ne.jp/

キーワードフィルタリング方式という技術を採用し，利用者がどの方式を利用するかを選択することができる．キーワードフィルタリングとは，サイト自体をブロックするのではなく，サイト内に現れる単語の中で不適切と思われる単語を「***」などの表示に置換え，有害情報を閲覧できないようにする方式である．しかし，これらのサービスもブラックリスト方式を主なフィルタリング方式としているので，先に挙げた問題点が依然として存在する．

2.2 有害サイトフィルタリングに関する研究

予め作成されたブラックリストに基づき，ブラックリストに登録されていない Web サイトにおいても，登録されている有害サイトとの類似性からフィルタリングを行う研究 [4] がある．この文献ではパターン認識手法を利用してブラックリストに登録されている文書との類似性を求めることによってフィルタリングを行っている．しかし，この手法はブラックリストを基準としてフィルタリングを行うため，人手によるリストの作成という問題点は解消されていない．本論文では，ブラックリストを作成することはせず，また，Web サイト単位でなく単語単位で有害確率を計算して有害サイトのフィルタリングを行う．

2.3 ベイジアンフィルタに関する研究

本研究では，ベイジアンフィルタに基づいた手法を用いて有害サイトのフィルタリングを行うが，ベイジアンフィルタの精度向上に関する研究は，迷惑メールのフィルタリング分野において多く行われている．迷惑メールのフィルタリングのためのベイジアンフィルタの研究は，文献 [5] 以後，広く行われるようになった．現在では，文献 [5] で提案されている PaulGraham 方式，PaulGraham 方式を改良した Robinson 方式 [6]，Robinson-Fisher 方式 [7] など，多くのフィルタリング方式が存在する．また，文献 [8] では，Bipolar 方式を提案している．これは，単語の有害確率が 0.5 から離れているものを 15 個抜き出して文書の有害確率を計算する PaulGraham 方式に対し，単語の有害確率が 1 に近いものと 0 に近いものを同数抽出することで，精度の向上や，スパム発信者側のフィルタ回避への対応を図っている．

このように，ベイジアンフィルタはその精度向上のための様々な研究がなされているが，語の共起による有害確率を考慮した研究は行われていない．

3 ベイジアンフィルタ

本章では，ベイジアンフィルタの代表的な方式である，PaulGraham 方式，Robinson 方式，Robinson-Fisher 方式について述べる．どの手法も以下の 4 手順で有害文書の判別を行うが，トークンの有害確率の計算方法および文書有害確率の計算方法がそれぞれ異なっている．なお，本論文における文書とは，Web 上に存在するページを指す．また，トークンとは文書を構成する最小単位であり，本論文においては単語をトークンとして用いている．

1. 文書 D からトークン w_i を取得する．
2. 全トークンの有害確率 $f(w_i)$ をそれぞれ計算する．
3. 文書 D に含まれるトークンの有害確率から文書 D の有害確率 $P(D)$ を計算する．
4. 文書の有害確率 $P(D)$ の値が設定された閾値以上であれば有害であると判別する．

3.1 PaulGraham 方式

文書から抽出したトークンごとに，トークンが有害文書に現れる確率 $p(w_i)$ を式 (1) で計算し，それをトークンの有害確率 $f(w_i)$ とする．

$$f(w_i) = p(w_i) = \frac{\frac{b_i}{n_{bad}}}{a \cdot \frac{g_i}{n_{good}} + \frac{b_i}{n_{bad}}} \tag{1}$$

ここで，a は非有害サイトの誤判定を防ぐためのバイアスであり，$a = 2$ とする．g_i，b_i は w_i が非有害ページと非有害ページにそれぞれ出現した回数，n_{bad} と n_{good} はそれぞれ有害ページと非有害ページの総数である．次に，文書 D に含まれるトークンの中で，有害確率 $f(w_i)$ と 0.5 の差の絶対値が大きい順に 15 個のトークンを抽出する．そして，それらの結合確率を式 (2) によって計算し，文書 D の有害確率 $P(D)$ とする．$P(D)$ が 0.7 以上であれば，文書 D は有害であると判定する．

$$P(D) = \frac{\prod_i f(w_i)}{\prod_i f(w_i) + \prod_i (1 - f(w_i))} \tag{2}$$

ここで，n は抽出したトークン数である．

3.2 Robinson 方式

Robinson 方式は，PaulGraham 方式を改良した方式であり，学習データ数が十分でない単語に対しての扱いを改善している．トークン w_i が有害文書に出現する確率 $p(w_i)$ は PaulGraham 方式と同様に式 (1) で計算する．（ただし，式 (1) のバイアス a を $a = 1$ として計算する．）そして有害確率 $p(w_i)$ からトークン w_i の有害確率 $f(w_i)$ を式 (3) で計算する．

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

x はトークンが一度も出現しなかったとき仮定として与える事前確率であり，s はその強度である．本研究ではそれぞれ $s = 1, x = 0.5$ とした．n_i はトークン w_i が出現した文書の総数であり，$n_i = g_i + b_i$ である．次に，文書 D の有害性 S と非有害性 H を次式で計算する．

$$S(D) = 1 - \{\prod_{i=1}^{n} (1 - f(w_i))\}^{\frac{1}{n}} \tag{4}$$

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

ここで n は文書 D に出現するトークンの総異なり数である．これらの値を用いて，文書が有害である確率 $P(D)$ を式 (6) で求める．そして $P(D)$ が 0.5 以上であれば文書 D は有害であると判定する．

$$P(D) = \frac{1 + \frac{H(D) - S(D)}{H(D) + S(D)}}{2} \tag{6}$$

3.3 Robinson-Fisher 方式

Robinson-Fisher 方式は Robinson 方式を改良した方式であり，トークンの有害確率の計算に Fisher の方法を用いている．トークンの有害確率 $f(w_i)$ は Robinson 方式と同様に式 (3) で計算される．文書 D の有害確率 $P(D)$ については，式 (7) と式 (8) で計算した有害性 $S(D)$ と非有害性 $H(D)$ から，式 (9) によって計算する．

$$S(D) = C(-2\ln\prod_{i=1}^{n} f(w_i), 2n) \tag{7}$$

$$H(D) = C(-2\ln\prod_{i=1}^{n} (1 - f(w_i)), 2n) \tag{8}$$

$$P(D) = \frac{1 - H(D) + S(D)}{2} \tag{9}$$

ここで n は文書 D に出現するトークンの総異なり数であり，$C(x, df)$ は自由度 df の χ^2 分布における x の片側確率である．そして，$P(D)$ が 0.5 以上であれば文書 D は有害であると判定する．この方式は，スパムフィルタである bsfilter[2]や，Thunderbird[3]などといったメールソフトで利用されている．

4 語の共起情報に基づく有害文書判別手法

4.1 概要

ベイジアンフィルタは，判別対象のページからトークンを抜き出し，過去の有害サイト，非有害サイトから得られた情報によってトークンの有害確率を計算する．そして，それらの結合確率によって，対象のページが有害かどうかを判定する．このとき，トークンの有害確率は 1 つ 1 つ独立に決定され，他のトークンの存在は影響しない．しかしそれでは，次のような問題が起きる．

例えば，出会い系サイトにおいては，「男性」，「女性」，「会員登録」，「出会い」，「ログイン」といった単語が頻出する．これらの単語をトークンとして抽出し，3 章で述べたような通常のベイジアンフィルタによって有害サイトかどうか判別しようとした場合を考える．「出会い」という単語は出会い系サイトに多く出現するため，有害確率は高くなると考えられるが，「男性」，「女性」，「会員登録」といった語は非有害サイトにおいても多く出現する一般的な語である．そのため，これらの語の有害確率は低くなり，結果としてこのサイトは有害ではないと誤判定されてしまうことが考えられる．

この問題を解決するために，本研究では 2 語の共起確率を利用する．「女性」，「出会い」といった単語の組が同一サイト上で共起した場合，そのサイトは女性との出会いを斡旋するサイトである確率が高くなる．よって，「女性」という単語も，そのサイトが有害サイトであることを説明する上で重要な要素となる．これは「会員登録」や「ログイン」といった語においてもいえることで，2 語の共起による有害確率の計算は，単語ベースのフィルタリングを行う上で有用であると考えられる．

以上の考え方に基づいて提案する有害文書判別手法の詳細を次節に示す．文書中に各トークンが出現した場合に，それらが有害文書に現れた確率からその文書が有害である確率を求めるというベイジアンフィルタの考え方に基づき，あるトークンが他のトークンと共起した場合に，そのトークンが有害となる確率を求める．

なお本研究では，文書中に出現した文を MeCab[4]を用いて形態素に分解し，そこから抽出した名詞と動詞をトークンとして用いる．

4.2 共起情報に基づくトークンの有害確率計算手順

本手法は，3 章で述べたベイジアン手法の 4 手順のうち，手順 2 のトークン w_i の有害確率 $f(w_i)$ の計算において語の共起情報を考慮する．トークン w_i の有害確率 $f(w_i)$ 計算以降の手順については，3 章で述べた各方式の計算方法をそのまま用いる．

1. **トークン w_i 単体での有害確率の計算**
   トークンの単独での有害確率 $f(w_i)$ は，既存の手法におけるトークンの有害確率と同様に，式 (1) や式 (3) によって求める．

2. **トークン w_i と他のトークンとの共起有害確率の計算**
   トークン w_i と w_j が有害文書で共起する確率 $p(w_i, w_j)$ を式 (10) で求め，その値を用いてトークンの共起有害確率を式 (11) で求める．

$$p(w_i, w_j) = \frac{\frac{cobad_{ij}}{b_i}}{a \cdot \frac{cogood_{ij}}{g_i} + \frac{cobad_{ij}}{b_i}} \tag{10}$$

$$f(w_i, w_j) = \frac{s \cdot x + n_{w_i} \cdot p(w_i, w_j)}{s + n_{w_i}} \tag{11}$$

   ここで b_i と g_i は式 (1) と，a，s，x は式 (3) とそれぞれ同じ値である．$cogood_{ij}$ と $cobad_{ij}$ は w_i と w_j が非有害ページ，有害ページで共起した回数，n_i はトークン w_i が出現した回数であり，$n_i = b_i + g_i$ である．

3. **トークンの有害確率の計算**
   トークン w_i との共起有害確率 $f(w_i, w_j)$ と 0.5 との差の絶対値が大きい順にトークン w_j を n 個抽出し，式 (12) と式 (13) でトークン w_i の有害性 $S(w_i)$ と非有害性 $H(w_i)$ を求める．

$$S(w_i) = C(-2\ln\{f(w_i)\prod_j f(w_i, w_j)\}, 2n) \tag{12}$$

$$H(w_i) = C(-2\ln\{(1 - f(w_i))\prod_j (1 - f(w_i, w_j))\}, 2n) \tag{13}$$

   $C(x, df)$ は自由度 df の χ^2 分布における x の片側確率である．また，抽出数 $n = 30$ とする．そして式 (14) でトークンの有害確率 $F(w_i)$ を計算し，この値をベイジアンフィルタの各方式で用いるトークンの有害確率とする．

$$F(w_i) = \frac{1 - H(w_i) + S(w_i)}{2} \tag{14}$$

[2]bsfilter : http://bsfilter.org/
[3]thunderbird : http://mozilla.jp/thunderbird/
[4]MeCab : http://mecab.sourceforge.net/

表 2: 評価実験 1 における Web ページの各カテゴリに対する誤判定率（単位：%）

方式	共起	有害サイト				非有害サイト			全体平均
		アダルト	出会い	風俗	有害平均	ニュース	ブログ	非有害平均	
PaulGraham	無	1.60	8.60	1.40	3.87	0.00	0.20	0.10	1.99
	有	0.40	1.60	1.00	1.00	0.00	0.40	0.20	0.60
Robinson	無	6.40	0.80	1.60	2.93	0.00	0.00	0.00	1.47
	有	5.20	0.20	1.40	2.27	0.00	2.40	1.20	1.74
Robinson-Fisher	無	0.00	0.20	0.00	0.07	0.00	0.80	0.40	0.24
	有	0.00	0.00	0.00	0.00	0.00	2.80	1.40	0.70

表 3: 評価実験 2 における Web ページの各カテゴリに対する誤判定率（単位：%）

方式	共起	有害サイト				非有害サイト				全体平均
		アダルト	出会い	風俗	有害平均	ニュース	ブログ	会員制	非有害平均	
Robinson-Fisher	無	0.40	1.20	0.20	0.60	0.00	1.00	1.80	0.93	0.77
	有	0.00	0.20	0.00	0.07	0.00	1.20	1.80	1.00	0.53

5 評価実験

PaulGraham 方式，Robinson 方式，Robinson-Fisher 方式の 3 方式について，それぞれ共起情報を用いない場合と用いた場合での比較評価を行う．5.1 節では，有害サイトから 3 カテゴリ，非有害サイトから 2 カテゴリをデータセットとして評価を行う．次に 5.2 節では，5.1 節で用いたデータに加えて，非有害サイトに「会員制サイト」を追加して同様の評価を行う．5.3 節では，学習データ数による性能の違いの評価を行う．以上のすべての評価は，以下に示す 5 分割交差検定によって行う．データセットを 5 等分し，各カテゴリを 100 件ずつ均等に組み合わせ 1 セットとする．そして，そこから 4 セットを学習データとし, 残りの 1 セットをテストデータとして評価を行う．

5.1 評価実験 1

5.1.1 データセット

有害サイトのデータとしては，表 1 に示されているカテゴリの中からアダルトサイト，出会い系サイト，風俗情報サイトを，非有害サイトとしてはブログとニュース記事をそれぞれ 500 件ずつ用いる．

5.1.2 結果

カテゴリごとの誤判定率を表 2 に示す．どの方式においても，共起情報を用いることで有害サイトの判別制度は向上していることがわかる．一方で非有害サイトに対しては誤判定率が上がってしまっているが，有害サイトフィルタリングにおいては，有害であるサイトを有害でないとする誤りはあってはならないため，有害サイトの判別精度が向上した本手法は有効であるといえる．

5.2 評価実験 2

5.1 節の評価では，学習データにおいて，非有害サイトとして 2 カテゴリしか用いていない．しかし実際には，非有害サイトにはより多くのカテゴリが存在しており，それらも学習データとする必要がある．

そこで，非有害サイトの種類を増やした場合の精度を評価するため，非有害サイトのデータとして会員制サイトを 500 件追加し再評価を行った．なお，評価にあたっては評価実験 1 で最も良好な結果であった Robinson-Fisher 方式を用いた．

5.2.1 データセット

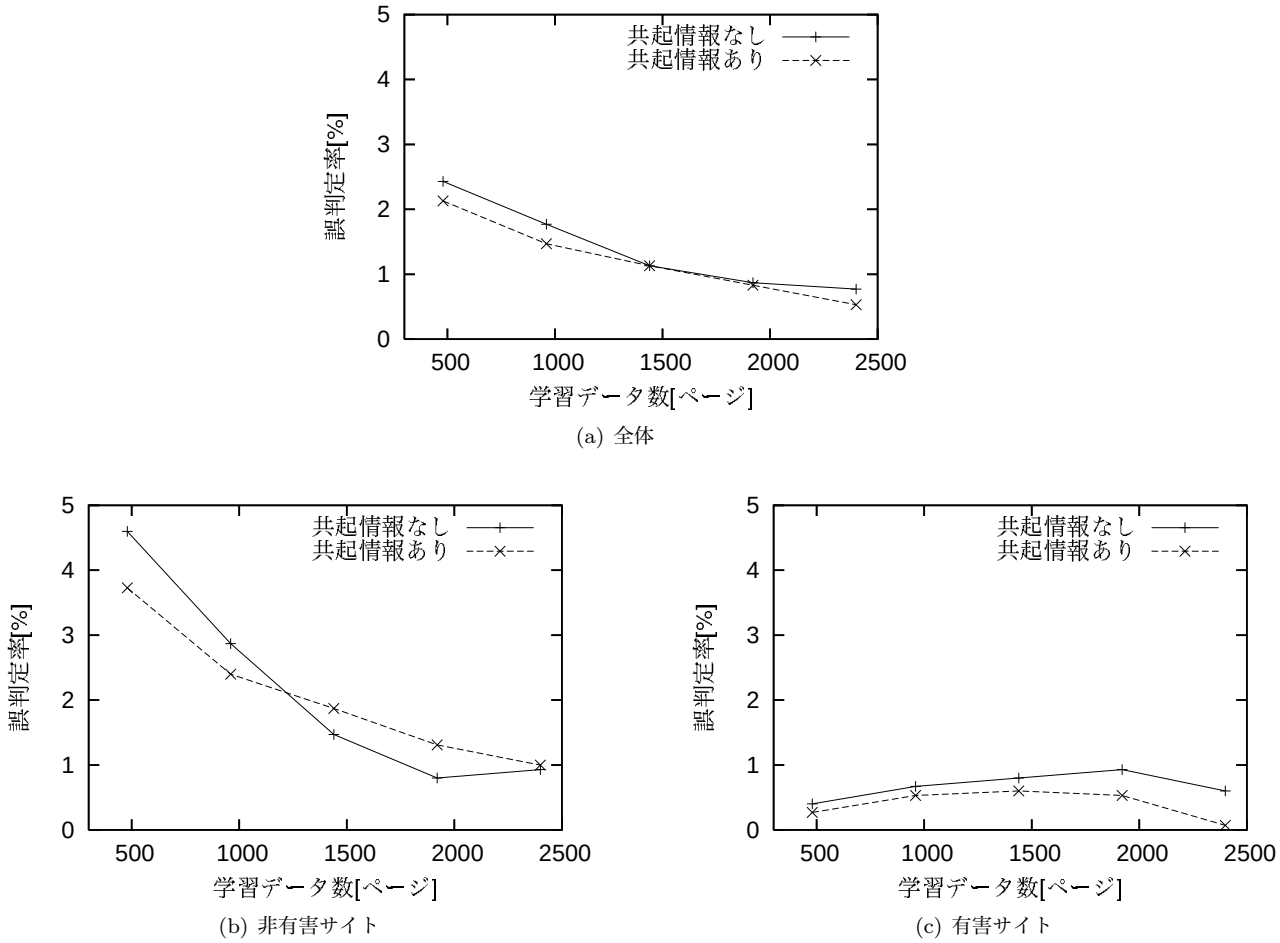
有害サイトのデータは評価実験 1 と同様で，非有害サイトのデータとして，ブログ，ニュース記事に加えて会員制サイトをそれぞれ 500 件ずつ用いた．

5.2.2 結果

各カテゴリに対する誤判定率を表 3 に示す．評価実験 1 では，Robinson-Fisher 方式は共起情報を用いない場合でも有害サイトの誤判定率は 0.07% と低い割合であり，共起情報を用いることでの精度向上は小さいように思えた．しかし，評価実験 2 において非有害サイトの学習データ数が増えたことで共起を用いない場合は有害サイトの誤判定率が 0.60% と増加しているが，共起を用いた場合では 0.07% であり，高い精度を保っている．

5.3 学習データ数と誤判定率の関係

評価実験 2 では，各カテゴリ 400 件ずつ，計 2400 件を用いて学習を行っている．ここでは，学習データ数を減らした場合の判別精度を評価した．全体の誤判定率を図 1(a)，有害サイトのみに対する誤判定率を図 1(b)，非有害サイトのみに対する誤判定率を図 1(c) にそれぞれ示す．図 1(a) からは，学習データ数を増やすことで順調に精度が向上していることがわかる．しかし，図 1(b) に示された有害サイトの誤判定率は学習データ数に比例して低下しているとはいえない．



(a) 全体

(b) 非有害サイト

(c) 有害サイト

図 1: 学習データ数による判別精度

5.4 有害確率の分布

本研究において，文書を有害とする閾値は，PaulGraham 方式では 0.7，Robinson 方式，Robinson-Fisher 方式では 0.5 と，既存手法に準じて設定した．しかし，共起情報を用いたことで，適切な閾値が変化している可能性は十分に考えられる．そこで，5.2 節における Robinson-Fisher 方式での評価において計算された文書の有害確率の分布を表 4 に示す．有害サイトでは極端な 1 件をのぞき，すべてのページが 0.50 を超えている．有害サイトフィルタリングにおいては，有害サイトを非有害であると誤判定してしまうことが最も問題となるので，今回設定した閾値 0.50 は適切であったと考えられる．

6 考察

6.1 共起情報利用による影響

共起情報を用いることで，ベイジアンフィルタのどの方式においても，有害サイトの判別精度を向上させることが可能であった．特に，実験 2 においては学習カテゴリ数が増えた場合に共起情報の有効性が増すということが示された．

しかし，評価実験 1 と同様に，評価実験 2 においても，有害サイトのカテゴリのうち，ブログに対して誤判定してしまう割合は共起情報を用いることで増えてしまっている．原因の 1 つとしては，実験に用いたデータセットにおいて，ブログのフォーマットで作成されている有害サイトが多かったことがあげられる．よって，ブログのフォーマットを構成する用語である，日付，時間,「コメント」や「トラックバック」などのブログに共通して出現する語との共起有害確率が，どの語においても高めに計算されてしまう傾向が出てしまい，結果として非有害ブログの有害確率が不当に高くなってしまったと考えられる．また，共起情報を利用する場合には，非有害サイトの学習数を多くすることが必要であるとも考えられ，この点は検討を要する．

6.2 学習データ数による精度の変化

5.3 節では，有害サイトにおいては，学習データ数と判別精度は比例関係にないという結果が得られた．この結果は，ただ単純に有害サイトと非有害サイトを多く学習すれば精度が向上するわけではないことを示している．今回の実験においては，学習データ数が多ければ多いほど全体の誤判定率は小さくなる結果となったが，ここからさらに学習データを増やした場合に，単純に精度が向上するとは限らない．今後より多くのデータを用いて実験を行い，学習データ数，学習比率が判定精度に与える影響を調査することは必要であると考えられる．

6.3 誤判定ページ

共起情報の使用の有無に関わらず誤判定されてしまったページには，Web サイトの入り口にあたるページが多く存在した．これらのページでは文章の量が少なく，Web サイトのメインページへのリンクが貼ってあるのみであるため，単語ベースでの有害性の判断は困難であったと考えられる．他にも，有害ではないと誤判定されてしまった有害サイト

表 4: Robinson-Fisher 方式における共起情報を用いた場合の有害確率 P(D) の分布

	範囲			
	$1.00 \geq P > 0.80$	$0.80 \geq P > 0.50$	$0.50 \geq P > 0.20$	$0.20 \geq P \geq 0.00$
有害サイト上	1489	10	0	1
非有害サイト上	5	10	38	1454

表 5: トークンの有害確率の例

トークン（単独確率）	共起語（共起有害確率）	有害確率
会員（0.43）	不倫（0.98）	0.99
	出会う（0.88）	
	恋人（0.96）	
登録（0.37）	出会い（0.97）	0.99
	ポイント（0.81）	
	近所（0.90）	
女性（0.61）	募集（0.94）	0.99
	アドレス（0.92）	
	メル友（0.99）	
男性（0.61）	アダルト（0.96）	0.99
	未満（0.93）	
	不倫（0.94）	
東京（0.51）	セフレ（0.99）	0.98
	出会い（0.90）	
	女の子（0.92）	

では，ページのごく一部のみが有害である情報を含むページなどもあった．これらを有害ページとするかは判断の分かれるところであるので，Yahoo!あんしんネットにおけるキーワードフィルタリングのような，有害と思われる箇所のみをフィルタリングする手法は有効であると考えられる．このようなフィルタリングを，単語の有害確率を考慮して行うことで，より柔軟なフィルタリングを行うことが可能になる．

6.4 共起有害確率

有害サイトである出会い系サイトにおいて計算されたトークンの有害確率の例を表 5 に示す．この表では，トークンの有害確率とともに，それらのトークンとの共起有害確率の高いトークンの例も示されている．表 5 において，「単独確率」は 3 章で述べた共起情報を利用しない手法で計算された有害確率 $f(w_i)$ を示しており，「有害確率」は共起情報を利用して本研究の手法で計算された確率 $F(w_i)$ である．

この表から，共起情報を用いることで，単独での有害確率が高くない語でも「出会い系サイトらしさ」を反映させた高い有害確率を求めるのが可能であることを示している．

7 おわりに

本研究では，語の共起情報に基づき，ベイジアンフィルタによって有害サイトのフィルタリングを行う手法を提案した．語の共起情報を用いることで有害サイトのフィルタリング精度が向上することが評価実験で確認され，単語の情報のみを用いた場合でも高精度なフィルタリングが可能であることを示した．一方で，共起情報により非有害サイトを有害であると判定してしまう割合が少しながら高くなってしまうことも明らかとなった．そのため，共起情報を用いる場合には，通常よりも非有害サイトの学習データが多く必要になるとも考えられる．今回の実験においては有害サイトと非有害サイトを 1：1 の割合で学習を行ったが，この比率および学習データ数の検討が必要である．

本研究においては，単語の情報のみを用いてフィルタリングを行ったが，Web ページ特有の情報である HTML タグの情報なども有効に活用することで，より判定精度を向上させることが可能であると考えられる．例えば，Web サイトの入り口ページなどを正しく判別するためには，ページのリンク先やリンク元の情報を用いることが効果的であると考えられる．

さらに，Web ページにおいては，ページ内に取り付けられた広告や，ブログのカレンダー，日付といった情報など，ページの内容とは本質的に関係のない情報も多く含まれている．それらの情報を排除した上で Web ページの内容の有害性を判定することも必要であり，Web ページの構造解析なども，単語ベースの Web フィルタリングにおいては重要であると考えられる．

参考文献

[1] 総務省: 通信利用動向調査, 報道発表資料, 平成 21 年 4 月 7 日 (2009).

[2] 財団法人インターネット協会: インターネット上の有害コンテンツの多様化に対応した新たな格付け基準 SafetyOnline3 の策定, 報道発表資料, 平成 19 年 4 月 3 日 (2007).

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

[4] 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.J84-D2, No.6, pp.1158-1166 (2001).

[5] P. Graham: A plan for spam, In P.Graham, *Hackers and Painters*, O'Reilly, pp.121-129 (2004).

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

[7] G. Robinson: A statistical approach to the spam problem, *Linux Journal*, No.107 (2003).

[8] 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, 情報処理学会研究報告, pp.73-76 (2007).