F-014

複数単語間の共起情報を用いた有害文書自動分類手法の提案
Filtering Harmful Sentences based on Multiple-Word Co-occurrence

藤井　雄太郎† 吉村　卓也‡ 伊藤孝行†‡§ 安藤　哲志£
Yutaro Fujii Takuya Yoshimura Takayuki Ito Satoshi Ando

1 はじめに

近年，掲示板やSNS(Social Network Service) のようなユーザーが自由に読み書きする事ができるWebサイトが増加している．また，Web上に存在する様々な情報の中にも，未成年に悪影響を及ぼす書き込みが存在し，問題となっている．これらの事実を受け，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] が行われており，有害な情報に対して対策をとる事が社会的に必要されてきている．多くのWebサイトでは，有害な情報に対しての対策をとっていない．また，対策を行っている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しているが，人による目視では，情報が発信されてから対処するまでの時間がかかる，情報量が膨大になった場合には処理が追いつかない等の問題が起こってしまう．そのため，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盛んに行われている．

既存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等が挙げられる．ブラックリスト方式では，有害な内容を含むWebサイトのURLをリスト化し，リストに挙げられたURLへのアクセスを制限する．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，リストに挙げられたURLのみを閲覧可能にする．ストップワード方式では，ブラックワードと称される単体で有害な意味を成す単語のリストを作成し，リストの単語が含まれるWebサイトへのアクセスを制限する．

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手法では，ブログやSNS全体にアクセス制限がかかってしまう，また個々のURLの設定を人手で行わなければならない事から，ユーザ数，用いられる単語の変化が著しい現在のWebサイトに対して適応する事は困難であると考えられる．ストップワード方式では，リストの作成に対して，ブラックワード選定基準等に関して大きなコストがかかるため，効率的ではなく，さらに隠語等が用いられる文書に対して対応できないため，やはり現在のWebサイトに適しているとは言い難い．

そこで本稿では，計算機を用いて文書の特徴を抽出し，自動的に有害，無害文書の分類を行う手法を提案する．具体的な手法として，過去のSNS及びブログに出現した有害文書と無害文書を，単語の共起の観点からそれぞれ計算機に学習させ，各文書の特徴を抽出し，文書の安全度を数値化する事で閾値と比較して分類を行う．本稿では，２単語間の共起情報，３単語間の共起情報を用いた手法を提案し，各手法及び既存の手法であるベイジアンフィルタと比較を行う事で，共起を用いる分類手法の有効性及び，共起の対象となる単語数に応じて精度の向上が見られる事を示す．さらに，複数のフィルタリング手法を組み合わせる事によって，精度の向上が見られる事も実験により検証する．本論文では，以下で，２章で関連研究について，３章では本論文において作成した共起辞書に構築方法と概要について述べる．続いて，４章では本論文での提案手法について，５章では提案手法における評価実験，考察，問題点の検証，及びブースティング[5]による用途の提案について述べ，６章でまとめる．

† 名古屋工業大学産業戦略工学専攻 School of Techno-Business Administration, Nagoya Institute of Technology
‡ 名古屋工業大学情報工学科 Information Engineering Department, 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 Toyo
£ NTT データ株式会社 NTT DATA Cooperation

2 関連研究

2 .1 既存の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また，Yahoo が提供するYahoo!あんしんネット[2] では，ブラックリスト方式，ホワイトリスト方式の2つに加えてキーワードフィルタリング方式を採用し，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存在する．

2 .2 サポートベクターマシン(SVM)

サポートベクターマシン(SVM)[3][4]は学習モデルの1つであり，機械学習の中で最も精度の良い手法の1つとして知られている．SVM は学習するデータの複数の特徴(素性)から分類できる超平面を求め，求めた超平面によってデータを分類する手法である．SVM では，超平面を求める際，境界に最も近いサンプルからのマージンを最大化することにより，未学習データへの精度を向上させている．しかし，多くの場合はノイズ等によって超平面で完全に分割し切ることができないため，ある程度誤りを許容するソフトマージンという手法が提案されている．SVM では，ベイジアンフィルタリングと違い，学習データが超平面を求める計算を再び行わなければならない．また，学習データが増加した場合，学習時間が急速に増加するという問題点がある．

2 .3 ベイジアンフィルタ

ベイジアンフィルタ[6][7]はスパムメールフィルタリングに使われる手法であり，スパムメール及び非スパムメー

ルから，各単語がスパムメールで出現する確率を学習し，メールに含まれるスパムメール，非スパムメールに特徴的な単語の出現確率の割合を計算することでフィルタリングを行う．

ベイジアンフィルタでは，学習の段階で正例(非スパム)と負例(スパム)を単語に分割し，各単語が正例，負例にそれぞれ何回出現したかを数え上げハッシュテーブルへと保存する．本稿では，作成したハッシュテーブルをデータベースへと保存し，利用している．作成したデータベースの構造を表１に示す．

表 1.　ベイジアンフィルタで用いるデータ構造

要素の説明	要素の型
WORD	STRING
w_iの出現回数(正例)	INT
w_iの出現回数(負例)	INT

フィルタリングの段階では，判定する文書を単語へ分割し，ハッシュテーブルから対応する単語の要素を取得する．取得した要素から単語w_iが正例に出現している確率P(w_i)を以下の式（１）で求める．ただし，b_i は単語w_iが正例に出現した回数，n_{good} は正例の総数，n_{bad} は負例の総数，aはバイアス変数である．

$$P(w_i) = \frac{\frac{(b_i)}{n_{good}}}{a * \frac{(b_i)}{n_{good}} + \frac{(g_i)}{n_{bad}}} \quad (1)$$

一般には，P(w_i)は負例(スパム)で出現する確率であるが，本稿では表現を提案手法と合わせる．P(w_i)は正例のし出現する確率とする．具体的には，判定する文書の各単語のP(w_i)から特徴的な単語を15単語W取得し，出現確率P(w_i)の総乗を求める．15単語は”A plan for spam[7]”によって経験的に決められている．ここで，特徴的な15単語は，P(w_i)の値が1．0に近い値から順に選ばれる．これはより特徴ある単語を抽出する事で，少ない単語数でも文書の特徴を抽出しやすくするためである．文書sentenceがスパムでない確率P(sentence)は，以下の式（２）で求める．ただし，w15∈sentenceはsentenceから取得したP(w_i)が特徴的な15単語の集合である．P(sentence)は[0．0-1．0]の実数を取り，0.0に近いほど有害文書の可能性が高く，1.0 に近いほど有害文書の可能性が低いと判定される．

$$P(sentence) = \frac{\prod_{W \in sentence} P(w_i)}{15} \quad (2)$$

また，ベイジアンフィルタでは，学習データが追加された場合，それまでに構築したハッシュテーブルを追加データ分だけ更新すれば良いため，新たなデータが追加された場合でも学習にかかる時間は追加データ分のみ計算すればいいという特徴がある．

3　共起辞書の構築

本稿では，統計的に共起情報を用いて有害文書の分類を行うため，共起情報を格納したデータベース(以下，共起辞書)を構築する．また，本稿における共起の定義として，３単語間の共起の場合，同一文書内に単語 w_1，w_2，w_3 が出現した時に，それらの組み合わせを共起関係 C_i(w_1，w_2，w_3)と定義する．２単語間の共起の場合も同様に C_i(w_1，w_2)，C_i(w_2，w_3)，C_i(w_1，w_3)のように定義する．また単語の出現順序は考慮しない．共起辞書には，Web 上から収集した正例と負例の各学習データを形態素解析ツールを用いて共起情報を抽出したデータを格納する．共起情報は共起の対象の単語数，学習量が増加すると，莫大な量になるため，計算機が処理しきれないという問題がある．そこで，本稿では，単語のハッシュ化，品詞の限定，Hadoop[8]を用いた分散処理により，本問題に対処する．以下の１．，２．，３．,４．，５．に共起辞書構築の一連の流れを示す．

【1.　学習データの収集】
【2.　学習データの正例、負例への分類】
【3.　文書から単語への分割】
【4.　単語，共起関係の ID 化(ハッシュ化)】
【5.　共起回数のカウント】

続いて，表 2，表 3，表 4，表 5，表 6 に本稿で構築した３単語間の共起辞書の概要を示す．

表 2.　単語ハッシュテーブル

要素の説明	要素の型
ID	INT
WORD	STRING

表 3.　２単語間共起ハッシュテーブル

要素の説明	要素の型
２単語間共起 ID	INT
WORD ID1	INT
WORD ID2	INT

表 4　２単語間共起頻度格納テーブル

要素の説明	要素の型
２単語間共起 ID	INT
C_iの出現回数(正例)	INT
C_iの出現回数(負例)	INT

表 5.　３単語間共起ハッシュテーブル

要素の説明	要素の型
３単語間共起 ID	INT
WORD ID1	INT
WORD ID2	INT
WORD ID3	INT

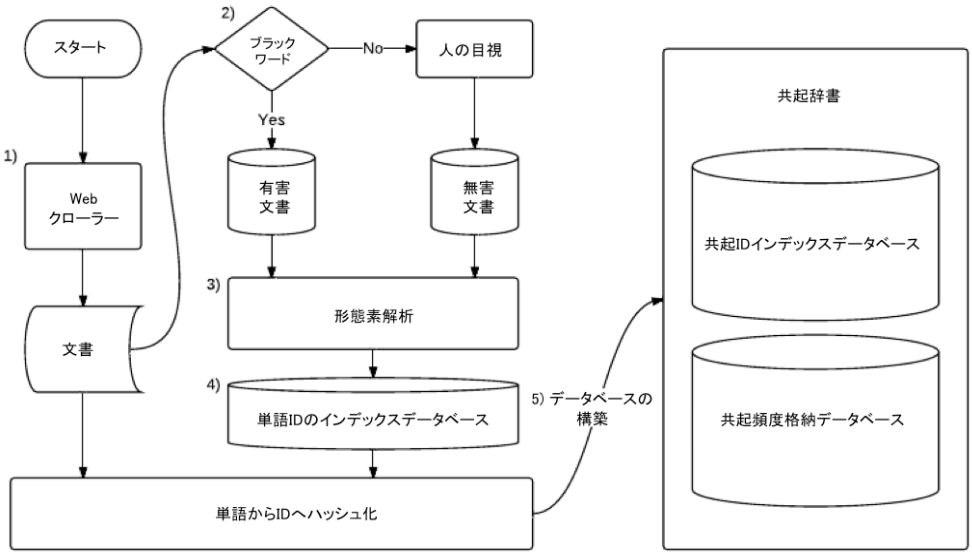

図 1.　共起辞書構築の概要図

表 6.　３単語間共起頻度格納テーブル

要素の説明	要素の型
３単語間共起 ID	INT
C_i の出現回数(正例)	INT
C_i の出現回数(負例)	INT

また，共起辞書の構築における概要図を図 1 に示す.

【 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自動的に収集する事で行う．本稿における収集データ件数は 23 万件である.

(正例に用いた文書の一例)
「今日は，授業中に携帯電話のバイブが鳴って，先生に怒られたー。皆にも笑われて、マジで今日は最悪だ。」

【 学習データの正例，負例への分類 】

収集した学習データを正例，負例にそれぞれ分類を行う．負例への分類においては，あらかじめブラックワードを収集しリスト化を行い，ブラックワードリスト内の単語を含む文書を負例と見なす事で，負例への分類を自動化する．本稿では，ブラックワードの数は 250 単語となっている．一方，正例への分類においては，人の目視による分類を行う．理由は，ブラックワードが含まれていない文書の中にも，隠語や比喩的な表現を含む文書が存在し，負例としてなり得るためである．また，本稿では正例，負例の学習量を統一するため，各学習データを正例と負例でそれぞれ 10 万件ずつ収集した．表 7 に学習データの内訳を示す.

表 7.　学習データの内訳

要素の説明	要素数
正例	100,000
負例	100,000
正例＋負例の総単語数	108,675

【 文書から単語への分割 】

文書から単語への分割を行うため，形態素解析ツール Mecab[9]を用いて，学習データを単語に分割する．また，分割する際に，品詞を限定して単語の抽出を行う．本稿では，Mecab の形態素解析により，品詞が助詞，助動詞と判定された単語は抽出しない．理由としては，共起の対象となる単語数を削減する事で計算量を削減し，単体で意味を成さない助詞，助動詞のような品詞ならば取り除いても精度に影響はないと考えられるからである.

【 単語・共起関係の ID 化(ハッシュ化) 】

計算量削減のため，単語及び共起関係の ID 化を行う．分割した単語の ID 化を行い(1)(2)(3)，構築した単語ハッシュテーブルの ID を用いて，共起関係の ID 化を行う(4).

(例)

(1) 文書	「今日は，良い天気だ。」	
(2) 単語への分割	{今日，良い，天気}	
(3) 単語の ID 化	[単語]	[単語 ID]
	「今日」 →	1
	「良い」 →	48
	「天気」 →	114
(4)共起関係の ID 化	[単語 ID の集合]	[共起 ID]
	{1,48,114} →	3

【共起回数のカウント】

ここでは，各共起の正例，負例における出現回数をデータベースに格納する．複数単語の共起を用るため，共起の総数が単語の冪乗オーダーとなり，学習データが増加した場合，計算機１台で共起をすべて数え上げることが難しくなる．そこで分散処理ツールである Hadoop を利用し，作成している．Hadoop は Google の MapReduce 及び GoogleFileSystem のオープンソース実装である．Hadoop は複数台の計算機で大規模なデータに対する処理を容易に行うことができるという特徴がある．また，Hadoop は Apache のトップレベルプロジェクトの１つであり，近年，盛んに開発，利用がされている．Hadoop では，OS のファイルシステムの上に HDFS(Hadoop Distribute File System)という独自の分散ファイルシステムを構築し，HDFS 上のファイルを用いて MapReduce を行う． Hadoop は Map，Shuffle，Sort，及び Reduce という 4 つのステップにより構成される．それぞれのステップでデータは分割され，複数が同時に実行される．Hadoop では，Map を行う Mapper と Shuffle，Sort，及び Reduce を行う Reducer が複数実行され，並列に分散して処理が行われる．Map は，ファイル中の各行(要素)を解析し，各行の解析結果を key,value の組み合わせで次のステップへ出力を行う．以降，key,value の組みを<key,value>と記述する．Shuffle では，Map の出力から各 Reducer への適切なデータ部分を取得する．Sort ステップでは，Shuffle で取得したデータを key 順にソートを行う．Reduce では，Sort によりソートされた Map の出力をまとめ，結果として HDFS 上へ出力を行う．

例えば，2 単語間の共起であれば，Map でのステップで，各共起を<[単語 A,単語 B],[0,1]>といったように，共起および共起が正例あるいは負例に１回出現したことを出力する．同じ行に同じ共起が複数もつ場合も同じように出力する．Shuffle によって同じ共起をもつ出力が同じ Reducer に集められ各共起が正例および負例にそれぞれ何度出現したかを数え上げ，各共起について<[単語 A,単語 B],[3,10]>といった結果を HDFS へ出力する．この場合，単語 A と単語 B は正例で３回，負例で 10 回共起している．

Hadoop には，Hadoop Streaming という拡張機能があり，Java 以外にも Perl ， Python や Ruby といった言語で MapReduce の記述が可能となっている．本稿では，Hadoop Streaming を Ruby を用いて実行している．正例及び負例 1 万件ずつで３単語間の共起情報を抽出した場合，Hadoop を使用せず作成すると 6〜7 時間程度掛かっていたが，計算機５台を用いて Hadoop を使用することで，1 時間程度に短縮することが可能となった．表 8 に構築した共起辞書の内訳を示す．

表 8.　共起辞書の内訳

要素の説明	要素数
２単語間の共起組み合わせ	53,310,776
３単語間の共起組み合わせ	1,498,732,588

4　複数単語間の共起情報を利用した有害文書分類手法の提案

本章では，２単語間の共起情報と３単語間の共起情報を用いた２つの手法を提案する．また，各手法においても，multiple と average の２つの計算方法を提案する．

4.1 2 単語間の共起情報を利用した有害文書分類手法

２単語間の共起情報を利用した有害文書分類手法では，同一文書 S 内に単語 w_1，と単語 w_2 が出現した時の共起 $C_i(w_1, w_2)$が有害でない確率 $P(w_1, w_2)$を求め，$P(w_1, w_2)$を用いて S が有害ではない確率を求める． $P(w_1, w_2)$は式(3)を用いて計算する．ここで，$b(w_1, w_2)$を，共起辞書内の $C_i(w_1, w_2)$の負例における出現回数とし，$g(w_1, w_2)$を共起辞書内の $C_i(w_1, w_2)$の正例における出現回数とする．また，$b(w_1, w_2)$，$g(w_1, w_2)$が共に 0 の場合にも計算を可能にするため，各要素に 1 を加算する．

$$P(w_1, w_2) = \frac{\{g(w_1, w_2) + 1\}}{\{g(w_1, w_2) + 1\} + \{b(w_1, w_2) + 1\}} \quad (3)$$

続いて，各 $P(w_1, w_2)$の総乗を用いて S が有害ではない確率を計算する multiple 手法と，各 $P(w_1, w_2)$の平均値を用いて S が有害ではない確率を計算する average 手法の２通りの計算方法で計算を行う．multiple 手法は式(4)を用いて計算する．

$$Safe\_2\_m(S) = \frac{\prod_{(w_1,w_2)\in S} P}{\prod_{(w_1,w_2)\in S} P + \prod_{(w_1,w_2)\in S} (1-P)} \quad (4)$$

average 手法は式(5)を用いて計算する．

$$Safe\_2\_a(S) = AVERAGE_{(w_1,w_2)\in S} P \quad (5)$$

ここで，Safe_2_m(S)を２単語間の共起情報を利用した有害文書分類手法で multiple 手法を用いて計算された S が有害ではない確率とし，Safe_2_a(S)を２単語間の共起情報を利用した有害文書分類手法で average 手法を用いて計算された S が有害ではない確率とする．S が有害ではない確率を閾値と比較し，閾値以下ならば，S を有害文書とし，閾値以上ならば，S を無害文書とする．

4．2　3 単語間の共起情報を利用した有害文書分類手法

３単語間の共起情報を利用した有害文書分類手法は，２単語間の共起情報を利用した有害文書分類手法の単語の共起数を拡張した手法である．同一文書 S 内に単語 w_1，単語 w_2，単語 w_3 が出現した時の共起 $C_i(w_1, w_2, w_3)$が有害でない確率 $P(w_1, w_2, w_3)$を求め，$P(w_1, w_2, w_3)$を用いて文書 S

が有害ではない確率を求める． P(w1, w2, w3)は式(6)を用いて計算する．ここで，b(w1, w2, w3)を，共起辞書内の Ci(w1, w2, w3)の負例における出現回数とし，g(w1, w2, w3)を共起辞書内の Ci(w1, w2, w3)の正例における出現回数とする．

$$P(w_1,w_2,w_3)=\frac{\{g(w_1,w_2,w_3)+1\}}{\{g(w_1,w_2,w_3)+1\}+\{b(w_1,w_2,w_3)+1\}} \quad (6)$$

続いて，各 P(w1, w2, w3)の総乗を用いて S が有害ではない確率を計算する multiple 手法と，各 P(w1, w2, w3)の平均値を用いて S が有害ではない確率を計算する average 手法の２通りの計算方法で計算を行う．multiple 手法は式(7)を用いて計算する．

$$Safe\_3\_m(S)=\frac{\prod_{(w_1,w_2,w_3)\in S} P}{\prod_{(w_1,w_2,w_3)\in S} P+\prod_{(w_1,w_2,w_3)\in S}(1-P)} \quad (7)$$

average 手法は式(8)を用いて計算する．

$$Safe\_3\_a(S)=AVERAGE_{(w_1,w_2,w_3)\in S}P \quad (8)$$

ここで，Safe_3_m(S)を３単語間の共起情報を利用した有害文書分類手法で multiple 手法を用いて計算された S が有害ではない確率とし，Safe_3_a(S)を３単語間の共起情報を利用した有害文書分類手法で average 手法を用いて計算された S が有害ではない確率とする．S が有害ではない確率を閾値と比較し，閾値以下ならば，S を有害文書とし，閾値以上ならば，S を無害文書とする．

5 評価実験

２単語間の共起情報を利用した有害文書分類手法，３単語間の共起情報を利用した有害文書分類手法とベイジアンフィルタ(PaulGraham 方式)の３方式についてそれぞれ比較実験を行う．全ての評価実験では，有害文書と無害文書のデータセットの分類を行い，実験における閾値は，0.5 に設定する．

1.1 multiple 手法における評価実験

本実験のテストデータとして，２ちゃんねるの掲示板から取得する．また，テストデータは文書内の単語数が 10 以上の文書とする． データ数として，有害文書 6500 件，無害文書 6500 件を用いる．評価実験は，収集したテストデータ正例，負例各 6500 件ずつを分類させ，再現率，適合率，及び F 値の観点から評価を行う．ベイジアンフィルタの分類結果を図２と表９に，２単語共起手法の分類結果を図３と表１０に，３単語共起手法の分類結果を図４と表１１に示す．また，各手法における再現率，適合率，F 値を表１２に示す．

表９． ベイジアンフィルタ実験結果

	「無害」分類	「有害」分類
無害文書	3105	3395
有害文書	6463	37

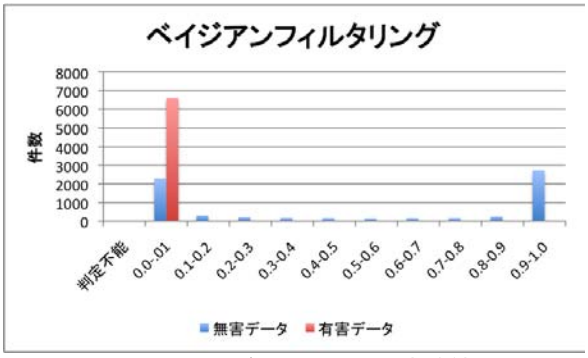

図 2． ベイジアンフィルタ実験結果

表 10． ２単語共起手法実験結果

	「無害」分類	「有害」分類
無害文書	3012	3488
有害文書	6554	46

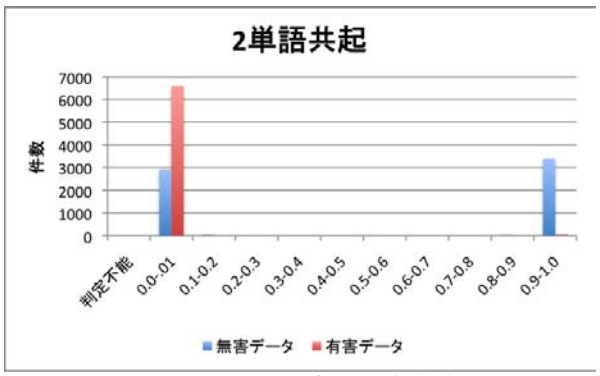

図３． ２単語共起手法実験結果

表 11． ３単語共起手法実験結果

	「無害」分類	「有害」分類
無害文書	2880	3620
有害文書	6458	92

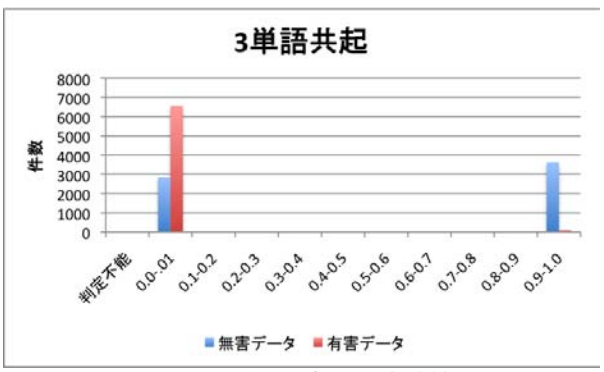

図 4． ３単語共起手法実験結果

表 12.　各手法における再現率，適合率，F 値

手法	データ種類	再現率	適合率	F 値
ベイジアンフィルタ	無害文書	52．21%	98．03%	0．6835
	有害文書	98．99%	68．49%	0．8071
2 単語共起手法	無害文書	55．57%	96．99%	0．6937
	有害文書	98．27%	69．28%	0．8097
3 単語共起手法	無害文書	52．21%	98．93%	0．7070
	有害文書	99．44%	67．92%	0．8128

図３，図４，図５，表９，表１０，及び表１１まで見てわかるように，値の分布，及び分類結果は類似した結果となっている．しかし，F 値に関しては，無害文書，有害文書の各テストデータの値もベイジアンフィルタよりも２単語共起手法の方が高く，また２単語共起手法よりも，３単語共起手法の方が高くなっている．以上の事から，共起を用いる事で，分類の精度が向上させる事ができた．また，共起の対象となる単語数が増加するに連れて，F 値が高くなっている事から，共起対象となる単語を増加させる事によって精度が向上する事も示す事ができたと考えられる．しかし，全ての手法において，無害文書の約半分の数が「有害文書」に誤分類されている事がわかる．この原因として，学習データに含まれる単語数の差が関係している可能性がある．

5．2　average 手法における比較実験

本実験のベイジアンフィルタの計算方法において，average 手法はベイジアンフィルタ本来の手法とは異なるため，本稿ではベイジアンフィルタ(ave)とし，また Safe(S)は式(9)を用いて計算する．　ここで，Safe_b_a(S)をベイジアンフィルタを用いて計算された S が有害ではない確率とする．また，P は式(1)の値を用いる．　対象となる単語は，文書内の全ての単語とする．

$$Safe\_b\_a(S) = AVERAGE_{w_i \in S} P \qquad (9)$$

本実験では，テストデータとして，２ちゃんねるの掲示板から取得する．　また，テストデータは文書内の単語数が 10 以上の文書とする．データ数として，有害文書 10,000 件，無害文書 10,000 件を用いる．評価実験は，収集したテストデータ正例，負例各 10000 件ずつを分類させ，分類精度の観点から評価を行う．ベイジアンフィルタの分類結果を図 6 と表 13 に，２単語共起手法の分類結果を図 7 と表 14 に，３単語共起手法の分類結果を図 9 と表 15 に示す．

表 13.　ベイジアンフィルタ(ave)実験結果

	「無害」分類	「有害」分類
無害文書	7549	2451
有害文書	9950	50

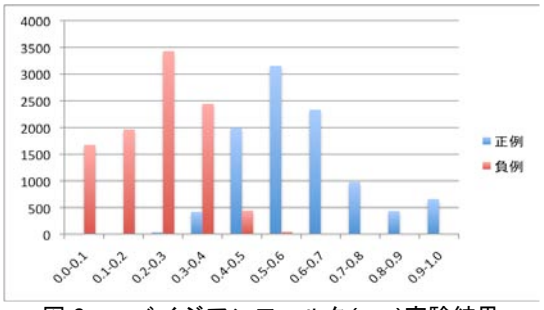

図 6.　ベイジアンフィルタ(ave)実験結果

表 14.　2 単語共起手法実験結果

	「無害」分類	「有害」分類
無害文書	9087	913
有害文書	9994	9994

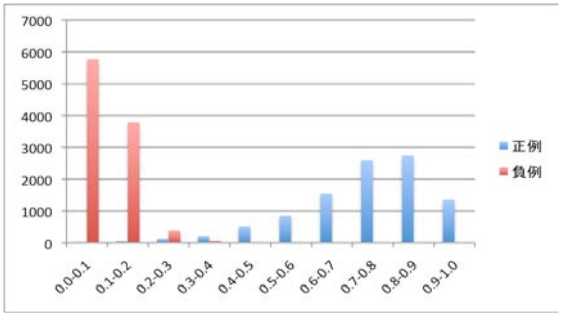

図 7.　2 単語共起手法実験結果

表 15.　3 単語共起手法実験結果

	「無害」分類	「有害」分類
無害文書	9857	143
有害文書	9999	1

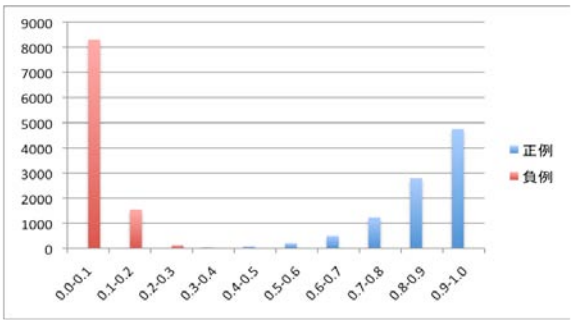

図 8.　3 単語共起手法実験結果

各手法の結果を比較すると，誤分類された文書の量が，ベイジアンフィルタ(ave)よりも，２単語共起手法，３単語共起手法の方が低く，また，２単語共起手法よりも３単語共起手法共起の方が低い結果となっている．これらの事から,共起の対象となる単語数を増加させると，精度が向上する事がわかるため，文書分類における共起の有効性を示す事ができると考えられる．

5．3　考察

multiple 手法の無害文書の誤分類

multiple 手法において，全ての手法で無害文書の約半分が「有害文書」に誤分類されていた．この原因として，本稿で構築した辞書の内容に原因があると考えられる．構築した共起辞書の正例と負例の数は等しいが，各学習データの内容を見ると,正例の文書に含まれる単語数の平均が約 20 単語に対して，負例の文書に含まれる単語数の平気が約 30 単語となっている事から，分類精度は学習データの内容によって大きく変化すると考えられる．

以上の事について，検証するため，新たに追加実験を行う．実験方法は，ベイジアンフィルタ(PaulGraham 方式)を用いて，正例，負例に含まれる単語数の比を変化させ，各単語数の比毎で分類結果を比較する．単語数の比は，正例：負例＝１：２，２：１，１：４，１：１の４つのパターンで行う．また，今回のテストデータは 5．2 節で用いたデータセットを用いる．

表 16 に単語比(正例：負例＝２：１)の実験結果，表 17 に単語比(正例：負例＝１：２)の実験結果，表 18 に単語比(正例：負例＝４：１)の実験結果，表 19 に単語比(正例：負例＝１：１)の実験結果を示す．

表 16.　単語数の比(正例：負例＝２：１)

２：１	「無害」分類	「有害」分類
無害文書	4427	3573
有害文書	7941	59

表 17.　単語数の比(正例：負例＝１：２)

１：２	「無害」分類	「有害」分類
無害文書	77	7923
有害文書	3892	4108

表 18.　単語数の比(正例：負例＝４：１)

４：１	「無害」分類	「有害」分類
無害文書	7221	779
有害文書	7997	3

表 19.　単語数の比(正例：負例＝１：１)

１：１	「無害」分類	「有害」分類
無害文書	523	7477
有害文書	6258	1742

表 16 では，単語数の比が正例の方が大きく，「無害」に分類されている文書も多くなっている．表 17 では，単語数の比が負例の方が多く，「有害」に分類された文書も多くなっている．表 18 では，表 16 よりもさらに正例の単語数の比を大きくしており，「無害」に分類された文書も，表 16 の結果よりも多くなっている．また，表 19 では，正例と負例の単語数の比を等しくした場合に，「有害」に分類された文書，「無害」に分類された文書ともに類似した数になっている．これらの結果から，各学習データの単語数の比は，精度に強く関係している事がわかった．今後は，単語数の比の差も考慮した計算方法を考案する必要がある．

average 手法と multiple 手法の比較

average 手法と multiple 手法の Safe(S)の値の分布を比較してみると,average 手法では一様に分布しており，multiple 手法では，両極端に分布している．これは multiple 手法の計算方法が総乗を用いている事が原因として考えられる．１つでも低い P の値が存在した場合，その文書の S が有害ではない確率の値まで 0 に収束してしまうからである．これらの比較から，average 手法の方が，より各文書毎にどの程度有害文書の確率が高いかという特徴を抽出していると考えられる．しかし，一方で閾値の設定に関しては，average 手法は，S が有害ではない確率の値が一様に分布しているために閾値の設定によって結果が大きく変化してしまう．そのために，average 手法に関しては，適切な閾値の設定方法が必要となる．

適切な学習量

今回 multiple 手法に関して，共起を用いた手法とベイジアンフィルタの間に大きな差が見られなかった．この原因として，各手法における適切な学習に差がある事が考えられる．各手法での単語数,もしくは共起の組み合わせの数は共起単語数の増加に比例して増加する．つまり，適合するパターン数が増加した分，学習量も増加する必要があると考えられる．

5．5　ブースティングによる精度向上

本節では，multiple 手法における評価実験の結果から，ベイジアンフィルタ，２単語共起手法，３単語共起手法の各手法のブースティングによる評価実験を行い，本稿で提案する手法がブースティングに適応する事を示す．また，テストデータは average 手法における評価実験で用いたデータセットを用いる．ブースティングでは，各手法において，どれか１つの手法でも「無害文書」と分類された文書は，全て「無害文書」に分類する．また，本実験における閾値は 0.5 の値に設定する．

表 20 に，各手法の組み合わせに対する再現率，適合率，F 値を示す．表 20 を見ると，どの手法も単体の時よりも F 値が高くなっている事がわかる．また，３つの手法を組み合わせた時が最も高い F 値を表している事から，組み合わせる手法の数を増やすと，精度も向上する事がわかる．これらの結果から，本稿で提案した手法は，ブースティングを用いた手法に対しても有効であり，さらに精度も向上する事が示せたと考えられる．

表 20.各手法の組み合わせにおける再現率，適合率，F 値

手法の組み合わせ	データ種類	再現率	適合率	F 値
ベイジアンフィルタ＋２単語共起手法	無害文書	57．76%	97．92%	0．7231
	有害文書	98．80%	70．40%	0．8220
ベイジアンフィルタ＋３単語共起手法	無害文書	61．12%	96．99%	0．7487
	有害文書	98．11%	71．94%	0．8301
２単語共起手法＋３単語共起手法	無害文書	57．64%	96．99%	0．7231
	有害文書	98．24%	70．28%	0．8293
ベイジアンフィルタ＋２単語共起手法＋３単語共起手法	無害文書	61．21%	96．91%	0．7503
	有害文書	98．05%	70．20%	0．8306

6　まとめと今後の課題

本稿では，複数単語間の共起情報を統計的に抽出する事で，共起辞書を構築し，その際には膨大な計算量を削減するため，Hadoop を用いた分散処理等を活用した．さらに構築した共起辞書から２単語共起を用いた手法と３単語共起を用いた手法の２つの手法を提案した．評価実験ではベイジアンフィルタの分類精度と比べ，共起を用いた手法の精度の方が高い事がわかり，さらに３単語共起手法が最も精度が高かった．以上の事から，文書分類において，共起を用いる事の有効性，共起の対象となる単語数を増加させる事による精度向上の特性を示した．また，計算方法を変える事により，結果の分布が変化する事も示した．特に average 手法による３単語共起の実験結果は高い分類精度を示す事ができた．単語数の比に比例して分類結果が変化する事に対する考察を行い，検証実験により，考察の正当性を示した．また，ブースティングによる検証実験では，提案した手法及びベイジアンフィルタを組み合わせる事により，精度が向上する結果を示し，新たな手法の用途を示した．

検証実験によって明らかになった分類精度は単語数の比に依存するという事に関して，ベイジアンフィルタの Robinson 方式等を参考にして，単語数の比を考慮に入れた文書の有害確率の計算方法の検討を行っていかなければならない．　また，multiple 手法における評価実験では，ベイジアンフィルタと２単語共起手法，３単語共起手法に大きな違いは見られなかった事からも，各手法における適切な学習量の調査を行っていかなければならない．　また，学習量が少ない環境化においても，精度の高い分類を行う事ができる手法も SVM 等を考慮しながら同時に検討する必要がある．　なぜならば，多くの学習量が必要になれば，収集に対するコストも増加してしまい，さらには計算量の観点から見ても，負荷が大きくなってしまうからである．さらに本稿では，単語の出現頻度のみを考慮した手法を提案したが，順序を考慮する事により，詳細な文書の特徴を抽出できる手法を検討していかなければならない．

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/s-news/2007/071210_4．html

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

[3] H．Drucker，C．Wu and V．Vapnik，”Support VectorMachines for Spam Categorization．”IEEE Trans．OnNeutral Networks，vol．10，no 5，pp．1048-1054，1999

[4] 小林大祐，松村真宏，木戸冬子，石塚満 ”知識検索サイトにおける不適切な投稿の分類”，情報処理学会全国大会講演論文集，vol．69，no2 pp2．561-2．562，2007

[5] 村田，家内，竹ノ内：ブースティングと学習アルゴリズム，電子情報通信学会誌，Vol．88，No．9，PP．724-729（2005）

[6] Paul Graham ： Better Bayesian Filtering．Proceedings of the 2003 Spam Conference，2003

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

[8] Hadoop，http://hadoop．apache．org/

[9] MeCab，http://mecab．sourceforge．net/

[10] 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）．

[11] 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″，情報処理学会第 72 回全国大会，2010

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

[13] 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”，第 29 回情報理論とその応用シンポジウム予稿集，pp．613-616，2006

[14] 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）．

[15] 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”，第 9 回情報科学技術フォーラム（FIT2010）講演論文集（2010）

[16] 本田崇智，山本雅人，川村秀憲，大内東，“Web サイトの自動分類に向けた特徴分析とキーワード抽出に関する研究，”情報処理学会研究報告 ICS，no．78，pp．1-4，2005

[17] 池田和史，柳原正，松本一則，滝嶋康弘，”係り受け関係に基づく違法・有害情報の高精度検出方式の提案”，第 2 回データ工学と情報マネジメントに関するフォーラム（DEIM2010），2010