情報処理学会論文誌 Vol.54 No.2 571–584 (Feb. 2013)

共起関係の抽出範囲を考慮した有害情報フィルタリング手法

中村 健二[1] 田中 成典[2] 山本 雄平[3,a)] 安彦 智史[3]

受付日 2012年5月14日, 採録日 2012年11月2日

概要：インターネットには，青少年の健全な育成に不適切な有害情報が存在している．これらの情報を機械的に判定する様々な有害情報フィルタリングの研究が行われている．その中でも，単語間の共起に基づき抽出した特徴を用いて有害情報を判定する手法が着目されている．その多くは，特徴抽出の処理範囲であるウィンドウサイズをページ全体や文の係り受け関係などの一定範囲として用いている．しかし，投稿された文書の範囲は多様であることから，適切な単語の共起関係が取得できない場合がある．そのため，誤った単語の組合せが特徴として抽出され，有害情報の判定精度が低下するという問題がある．そこで，本研究では，ページ分割手法を用いて多様なウィンドウサイズを考慮した有害情報フィルタリング手法を提案する．そして，本提案手法の有用性を検証するため，既存手法との比較実験を実施した結果，本提案手法の方が高精度に判定可能であることを実証した．

キーワード：有害情報フィルタリング，Web ページ分割，共起関係

Method of Filtering Harmful Information Considering Extraction Range of Word Co-occurrence

Kenji Nakamura[1] Shigenori Tanaka[2] Yuhei Yamamoto[3,a)] Satoshi Abiko[3]

Received: May 14, 2012, Accepted: November 2, 2012

Abstract: The Internet contains a harmful information that is not conducive to the healthy development of young people. Researchers are investigating ways to mechanically identify this harmful information and to filter it. The methods that have received the most attention are many researches in which harmful information is identified using characteristics extracted based on the word co-occurrence. Many previous researches have been used the window size that is the scope of characteristic extraction, fixed lengths such as the whole-page and the dependency relationships. However, since the length of a submitted document is variable, a contradiction arises between the submitted document and the scope of the characteristics extraction. As such, incorrect word combinations are extracted as having a characteristic. This lowers the accuracy of harmful information identification. In this research, a method of filtering harmful information is proposed which accounts for variable window size using the page segmentation method. The utility of the proposed method was investigated by conducting that compared the method to previous ones. The results proofed that the proposed method has the potential for more accurate identification.

Keywords: Filtering harmful information, segmentation of Web page , Word co-occurrence

1. はじめに

携帯電話の利用者が低年齢化しており，携帯電話を利用してインターネットに接続する青少年が増加しつつある．2009 年 5 月に文部科学省が実施した「子どもの携帯電話等の利用に関する調査」[1] によると，小学 6 年生の 24.7%，中学 2 年生の 45.9%，高校 2 年生の 95.9%が携帯電話を保持しており，大多数の子どもは，携帯電話を用いてイン

1 大阪経済大学情報社会学部
Faculty of Information Technology and Social Sciences, Osaka University of Economics, Osaka 533–8533, Japan
2 関西大学総合情報学部
Faculty of Informatics, Kansai University, Osaka 569–1095, Japan
3 関西大学大学院総合情報学研究科
Graduate School of Informatics, Kansai University, Osaka 569–1095, Japan
a) yamamoto@kansai-labo.co.jp

© 2013 Information Processing Society of Japan

ターネットに接続している状況である．また，インターネットのコンテンツにも変化が生じており，SNS（Social Networking Service）や電子掲示板，ブログなどの CGM（Consumer Generated Media）の普及にともない，個々人が発信した多種多様な情報が大量に蓄積されている．しかし，これらの中には，援助交際に関わる情報などの青少年の健全育成に悪影響を及ぼす可能性のある情報（以下，有害情報）も含まれており，様々な問題が発生している．このための対応として，2009 年 4 月 1 日に「青少年が安全に安心してインターネットを利用できる環境の整備等に関する法律（青少年インターネット環境整備法）」[2] が施行され，有害情報フィルタリングの普及促進などにより青少年が有害情報を閲覧する機会を最小化する取り組みが実施されている．一般に利用されている有害情報フィルタリングサービスでは，ブラックリスト方式やホワイトリスト方式のフィルタリング手法が用いられている．これらの手法では，ドメイン単位や URL 単位での閲覧可否を設定可能であるが，人手でこれらのリストを作成，更新するコストが膨大となっており，今後も増加し続けるすべてのコンテンツに即応することは困難である．そのため，Web ページに含まれる情報から自動的に有害情報が含まれているかどうかを判定する研究 [3], [4], [5], [6], [7], [8], [9], [10], [11], [12] が実施されている．

有害情報を自動的に判定する手法は，大別して 2 つに分類される．1 つ目は，有害情報に関連するキーワードをあらかじめ用意し，そのキーワードに一致する割合に基づき対象の Web ページが有害であるかを判定するキーワード一致による手法 [3] である．2 つ目は判別対象とする有害情報の教師データを事前に準備し，その教師データから抽出した特徴に基づき有害情報を判定する教師あり学習による分類手法 [3], [4], [5], [6], [7], [8], [9], [10], [11], [12] である．

キーワード一致による手法 [3] では，適切なキーワードを設定することで，有害な情報を網羅的に収集できる．特に，援助交際などを示す隠語（「神待ち」や「円光」など）を用いて検索した際には，援助交際に関わるページが検索結果に多く含まれており，誤抽出は含まれるもののインターネット上の有害情報を網羅的に抽出できる．しかし，有害情報に関するキーワードは，時代に合わせて日々変化しており，つねに辞書のメンテナンスを行わなければ，時間経過とともに最新の情報をフィルタリングできない状況となる．そのための対応として，有害情報が含まれる文章情報を解析して，自動的に有害語辞書を構築する研究 [13], [14] が提案されている．

教師あり学習による分類手法 [3], [4], [5], [6], [7], [8], [9], [10], [11], [12] では，教師データを解析して有害情報を示す特徴を学習し，その学習した特徴に基づき未知の情報が有害情報である度合いを判別器により算出する．有害情報の判定に用いられる代表的な判別器とそれらを用いた手法として，SVM（Support Vector Machine）[15] を用いた手法 [3]，Naive Bayes Classifier [16] を用いた手法 [7]，Neural Network [17] を用いた手法 [5] などが提案されている．これらの研究では，有害情報を判別するための特徴データの定義がそれぞれ異なり，教師データに含まれる形態素を用いる手法 [3], [4], [5], [6], [7], [10]，形態素の組合せを用いる手法 [8], [11], [12], [14]，HTML のタグの出現頻度から算出した情報量を用いる手法 [6], [9] などが提案されている．

近年，新たに提案されている有害情報フィルタリングの手法 [8], [12], [14] では，単語間の共起に基づいて有害情報の特徴を抽出している．また，特徴抽出の処理範囲（以下，「ウィンドウサイズ」）として，ページ全体や文の係り受け関係が採用されている．しかし，インターネットの電子掲示板，ブログや SNS などは，Web ページによって文章の長さが一定ではなく，また，Web ページのデザインも Web サイトによって様々であり，Web ページ内に複数のウィンドウサイズが混在している状況である．そのため，従来のすべての Web ページに対して一定のウィンドウサイズを用いる手法では，適切な単語の共起情報を抽出できず，有害情報判定の精度低下につながる様々な課題が発生する．具体例として，ウィンドウサイズが小さい場合には，Web ページの文章が 1 つのウィンドウに収まらず適切な単語の組合せが抽出できないという問題やウィンドウ内に含まれる単語が少ないため共起関係抽出できないという問題などが発生する．一方，ウィンドウサイズが大きい場合には，メニューの単語とメインコンテンツの単語との共起関係が抽出され，不適切な共起語が抽出されるという問題や共起関係にある単語の組合せ数が膨大となり不要な共起語が抽出されるという問題などが発生する．

そこで，本研究では，ウィンドウサイズが一定であるという問題を解消するため，Web ページの見た目の特徴に基づきブロック単位に分割し，そのブロックをウィンドウサイズとする手法を提案する．そして，本提案手法を用いた有害情報フィルタリングの有用性を検証する．

2. 研究の概要

2.1 研究の目的

本研究では，有害情報フィルタリングのための特徴抽出の処理範囲であるウィンドウサイズが一定であることにより起因する問題の中でも，ウィンドウサイズを Web ページ全体とした場合に発生する「不適切な共起語が抽出されるという問題」と「共起関係にある単語の組合せ数が膨大となるという問題」を解消し，有害情報の判定精度を向上させることを目的とする．

そこで，Web ページをヘッダ，フッタ，メニュー，メインコンテンツのテキスト情報や表などの細分化された複数の領域（以下，ブロック）に分割し，そのブロックをウィンドウサイズとする手法を提案する．この手法により，各

© 2013 Information Processing Society of Japan

投稿記事やフッタなどをそれぞれブロックとして抽出できるため，共起関係の抽出範囲が限定され，不適切な共起語が抽出されるという問題を解決できると考えられる．

また，Web ページをブロック単位に分割することで抽出される共起語の数も削減できる．100 種類の単語が含まれる Web ページを 10 個のブロックに分割し，各ブロックに 10 種類ずつ単語が含まれた場合を例として共起語の組合せ数を試算すると，従来手法において 2 つ組の共起語，3 つ組の共起語数はそれぞれ 4,950 件（${}_{100}C_2$），161,700 件（${}_{100}C_3$）となるのに対して，本提案手法では 450 件（${}_{10}C_2 \cdot 10$），1,200 件（${}_{10}C_3 \cdot 10$）となり，抽出される共起語数を大幅に削減できることが分かる．このことから，共起関係にある単語の組合せ数が膨大となるという問題も解決できると考えられる．

2.2 ブロック分割手法の選定

Web ページをブロック単位に分割する手法としては，主として次に示す 2 つの手法が考えられる．

- Web ページ内のテキストの内容に基づき分割する手法
  本手法では，Web ページ内に含まれる特徴に基づき分割する．主として，テキストセグメンテーション [18] 技術などを用いて Web ページの文章を内容単位に分割する．
- Web ページを見た目に基づき分割する手法
  本手法では，Web ページの各 HTML 要素の画面上での位置に基づき，ブロック単位に分割 [19] する．たとえば，ヘッダ，フッタ，メニュー，メインコンテンツ，メインコンテンツ内の記事や画像など，Web ページのレイアウト構造に従って Web ページを分割する．

「Web ページ内のテキストの内容に基づき分割する手法」では，一般的に文章に着目して Web ページを分割する．そのため，メニュー部やヘッダ部などの主として単語や画像のみで構成される部分については，正しくブロックに分割できないと考えられる．一方，「Web ページを見た目に基づき分割する手法」では，Web ページの見た目に基づき分割するため，メニュー部や，メインコンテンツ部など，Web ページのデザインに従って複数のブロックに分割できる．このことにより，メインコンテンツ以外に含まれる Web ページの特徴も正しく抽出できると考えられる．そのため，本研究では，「Web ページを見た目に基づき分割する手法」を用いて Web ページをブロック単位に分割する手法を提案する．

2.3 処理の流れ

本提案手法は，**図 1** に示すとおり事前処理部と判定処理部がある．事前処理部は，Web ページ分割機能，共起検出機能と有害判定確率辞書の構築機能の 3 つの機能で構成される．判定処理部は，事前処理部で出力した結果を用いて

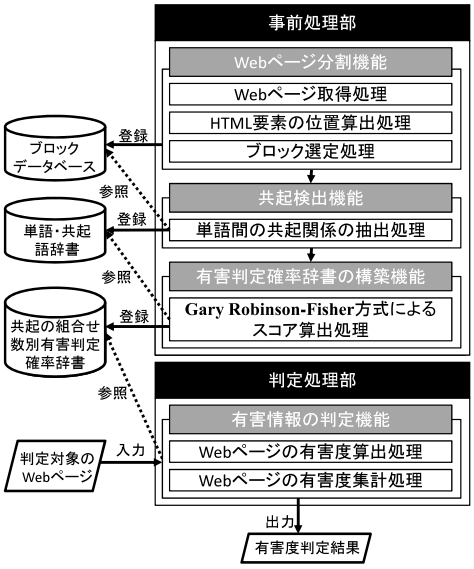

図 1 処理の流れ
Fig. 1 Flow of process.

判定対象の Web ページの有害度判定を行う有害情報の判定機能で構成される．なお，本研究では共起語を用いた有害情報の判定手法として一般的 [8], [12] に用いられる Gary Robinson-Fisher 方式 [20] を採用する．各処理の手順を次に示す．事前処理部の Web ページ分割機能では，有害情報判定の教師データとなる Web ページをブロック単位に分割する．そして，分割したブロックを Web ページの分割結果として，ブロックデータベースに登録する．共起検出機能では，まず，ブロックデータベースに含まれる文章を形態素解析して単語を抽出する．次に，抽出した単語の組合せである共起語を生成する．そして，単語，共起語の組合せを単語・共起語辞書に登録する．なお，形態素解析器には MeCab [21] を使用する．有害判定確率辞書の構築機能では，単語・共起語辞書に含まれる語句が出現する文章の傾向に基づき Gary Robinson-Fisher 方式で，語句の有害度を算出する．そして，算出した結果を共起の組合せ数別有害判定確率辞書に登録する．判定処理部の有害情報の判定機能では，有害判定確率辞書の構築機能で構築した共起の組合せ数別有害判定確率辞書を用いて，入力された判定対象の Web ページの有害度を算出する．

3. 学習アルゴリズム

3.1 Web ページ分割機能

本研究では，Web ページに記述された内容から適切に単語の共起関係を抽出するため，教師データの Web ページを見た目に基づきブロック単位に分割する．本研究では，ブロック単位に分割された中から，共起関係の抽出対象となるブロックを選定する必要があるため，HTML 要素の包含関係に基づき取捨選択する．Web ページ分割のイメージを**図 2** に示す．

© 2013 Information Processing Society of Japan

Algorithm 1　Web ページ分割アルゴリズム

Require:
- $ElementArea(n)$ //レイアウト要素 n の要素面積を算出
- $MaxElementArea(N_i)$ //レイアウト要素集合 N_i の中で最大の $ElementArea(n)$ を算出
- $Height(n)$ //レイアウト要素 n の高さを算出
- $Width(n)$ //レイアウト要素 n の幅を算出
- $ChildElements(n)$ //レイアウト要素 n の子要素を抽出
- $ParentElement(n)$ //レイアウト要素 n の親要素を抽出

Ensure:
- **Function** $NodeSelection(N_i)$
  - $N'_i := \{\}$ //初期化
  - //要素面積が 0 より大きく，$MaxElementArea(N_i)$ の閾値 α 未満となる n_{ij} を抽出
  - **While** $j < Length(N_i)$ **Do**
    - **If** $ElementArea(n_{ij}) > 0$
    - **Or** $ElementArea(n_{ij}) < MaxElementArea(N_i) \cdot \alpha$ **Then**
      - $N'_i := N'_i + n_{ij}$ //条件に一致する n_{ij} を N'_i に追加
    - **End If**
  - **End While**
  - //子要素と兄弟要素が 0 個のレイアウト要素を除外
  - **While** $k < Length(N'_i)$ **Do**
    - **If** $Count(ChildElements(n_{ik})) = 0$
    - **And** $Count(ChildElements(ParentElement(n_{ik}))) = 1$ **Then**
      - $N'_i := N'_i - n_{ik}$ //条件に一致する n_{ij} を N'_i から除外
    - **End If**
  - **End While**
  - //子要素の合計面積が自身の面積の閾値 β 以上を占めるレイアウト要素を除外
  - **While** $l < Length(N'_i)$ **Do**
    - **If** $Sum(ElementArea(ChildElements(n_{il})))/ElementArea(n_{il}) >= \beta$ **Then**
      - $N'_i := N'_i - n_{il}$ //条件に一致する n_{il} を N'_i から除外
    - **End If**
  - **End While**
  - $B_i := N'_i$ //B_i を作成
  - **Return** B_i
- **End Function**

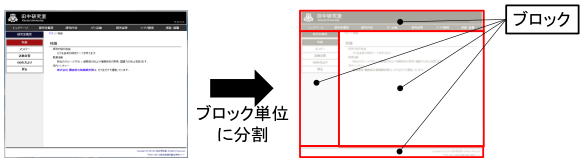

図 2　Web ページの分割例
Fig. 2　Example of segmentation of Web page.

本提案手法では，単語の出現位置の共通性を考慮するため，Web ブラウザの機能により構築された DOM（Document Object Model）から HTML 要素を取得し，それぞれの位置と要素面積の包含関係に基づきグループ化する．DOM を用いることで，Web ブラウザ上に表示するすべての HTML 要素と記述内容が取得できるため，CSS（Cascading Style Sheets）を用いてレイアウトが定義されている Web ページやスクリプト言語などを用いて動的にレイアウトが構築される Web ページなどにも対応可能である．具体例として，ユーザ操作により動的にレイアウトが構成されるアコーディオン UI などにおいても，ユーザ操作で表示される HTML 要素が DOM に含まれていれば要素面積が 0 の HTML 要素として取得できるため，包含関係を取得できる．そして，HTML 要素の位置と要素面積の包含関係に基づきグループ化した HTML 要素をブロックとして抽出する．Web ページ分割手法の詳細を **Algorithm 1** に示す．

本アルゴリズムは，Web ページ集合 $\{P_1, \ldots, P_i\}$ に含まれる任意の Web ページ P_i のすべての HTML 要素からノイズとなるタグを除去したレイアウト要素集合 $N_i = \{n_{i1}, \ldots, n_{ij}\}$ を解析し，ブロックとなるレイアウト要素のみを抽出したブロック集合 $B_i = \{b_{i1}, \ldots, b_{im}\}$ を作成する処理である．Algorithm 1 において，N_i は，P_i に含まれるすべてのレイアウト要素集合，B_i は P_i に含まれるすべてのレイアウト要素 N_i の中でブロックとして認識した n_{ij} からなるブロック集合を指す．なお，Algorithm 1 における親子関係とは，Web ページの表示内容の包含関係（**図 3**）を指す．本研究では，包含関係の算出に表示内容

© 2013 Information Processing Society of Japan

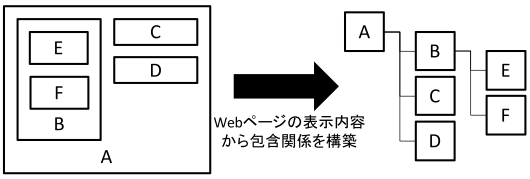

図 3 レイアウト要素の包含関係
Fig. 3 Inclusive relation of layout elements.

であるレイアウト要素の面積を用いる．レイアウト要素の面積は，レイアウト要素の高さと幅のピクセル数から算出する．

本処理では，次に示す 4 つのステップを順次実行することにより，ブロックとなるレイアウト要素を抽出する．なお，<input> タグ，<script> タグ，
 タグ，<hr> タグ，<iframe> タグ，<font> タグ，<b> タグについては，ユーザ入力のタグ，内部にテキストを持たないタグ，または文章の一部に多く使われるタグであり，ブロック分割時の精度低下につながると考えたため，ノイズとして事前処理で除去する．

STEP 1. Web ページ P_i に含まれるすべてのレイアウト要素集合 N_i から，次の 2 つの条件に一致するレイアウト要素 n_{ij} を抽出し，新たなレイアウト要素集合 N'_i に追加する．1 つ目の条件は，要素面積 $ElementArea(n_{ij})$ が 0 以上となることである．これは，画面上に表示されているレイアウト要素のみを選択するためである．2 つ目の条件は，要素面積 $ElementArea(n_{ij})$ が N_i の中で最も大きい要素面積である $Max(ElementArea(N_i))\cdot\alpha$ $(0 \leqq \alpha \leqq 1)$ 未満となることである．これは，<body> タグ直下に タグや <table> タグなど，ページ全体を被うスタイルが複数定義されている場合，ブロックがそのタグのみとなり，適切な処理結果を得ることができないためである．

STEP 2. STEP 1 で構築したレイアウト要素集合 N'_i から，次の 2 つの条件に一致するレイアウト要素 n_{ij} を除外する．1 つ目の条件は，レイアウト要素数 $Count(ChildElements(n_{ik}))$ が 0 であることである．これは，子要素数が 0 のレイアウト要素は，末端の要素でありレイアウトを構成するブロックとして不適切であると考えたためである．2 つ目の条件は，レイアウト要素 n_{ij} と親要素が同一である兄弟レイアウト要素数 $Count(ChildElements(ParentElement(n_{ik})))$ が 1 となることである．これは，兄弟要素が存在しないため親要素に含めることが適切であると考えたためである．

STEP 3. STEP 2 で不要な要素を除外したレイアウト要素集合 N'_i から，各レイアウト要素が持つ子要素の合計面積 $Sum(ElementArea(ChildElements(n_{il})))$ の占める割合を算出する．そして，算出した面積の割合がレイアウト要素面積 $ElementArea(n_{il})$ に対して閾値 β $(0 \leqq \beta \leqq 1)$ 以上となる n_{il} を N'_i から除外する．これは，子要素の占める割合が自身の領域の一定値を超える場合，子要素それぞれを独立したブロックとして抽出するためである．

STEP 4. STEP 3 の処理終了時に残った N'_i をブロック集合 B_i とする．

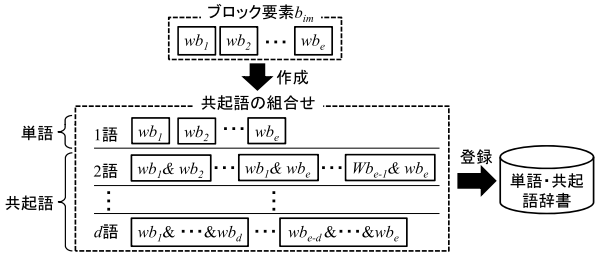

図 4 共起関係の抽出手法
Fig. 4 Extraction method of word co-occurrence.

3.2 共起検出機能

本提案手法による共起関係の抽出手法のイメージを図 4 に示す．本機能では，まず，Web ページ分割機能で得られたブロック集合 B_i の各ブロックに含まれる文章を MeCab で形態素解析する．次に，各ブロック b_{im} において，単語が同時に登場している場合，共起語の組合せを作成する．共起語の組合せは，単語のみも含むため 1 語から d 語までの組合せとし，共起関係が抽出できる対象すべてを抽出する．最後に，抽出した各共起語群 $W = w_1 \cup w_2 \cup \cdots \cup w_d$ に含まれる各共起語 $w_d = \{wb_{d1}, \ldots, wb_{dq}\}$ を単語・共起語辞書に登録する．

3.3 有害判定確率辞書の構築機能

本機能では，単語・共起語辞書に登録された各語句の有害度を算出し，共起語の組合せ数別有害判定確率辞書を構築する．本研究では，有害度の算出に有害判定手法として一般的に利用される Gary Robinson-Fisher 方式を用いる．共起語 wb_{dq} の有害度 $p(wb_{dq})$ は，式 (1) で算出する．なお，有害度の算出には，有害な Web ページから抽出した有害ブロック集合 BH と無害な Web ページから抽出した無害ブロック集合 BS を用いる．

$$p(wb_{dq}) = \frac{n_{BH}(wb_{dq})/n_{BH}}{n_{BH}(wb_{dq})/n_{BH} + n_{BS}(wb_{dq})/n_{BS}} \quad (1)$$

また，$n_{BH}(wb_{dq})$ は共起語 wb_{dq} が出現する有害ブロックの数を表し，$n_{BS}(wb_{dq})$ は wb_{dq} が出現する無害ブロックの数を表す．wb_{dq} の有害度 $p(wb_{dq})$ が 0 に近いほど，wb_{dq} は無害ブロックにおける特徴的な共起語であることを意味し，$p(wb_{dq})$ が 1 に近いほど，wb_{dq} は有害ブロックにおける特徴的な共起語であることを意味する．式 (1) では，1 件のブロックに wb_{dq} が出現する確率を比較することで，有

© 2013 Information Processing Society of Japan

害ブロック数と無害ブロック数の差に影響されずに共起語の有害度を算出できる．しかし，共起語 wb_{dq} の出現回数が少ない場合，適切な有害度を算出できないという問題がある．そこで，Gary Robinson-Fisher 方式では，$p(wb_{dq})$ を上記の問題に対応できるように補正している．補正後の wb_{dq} の有害度 $f(wb_{dq})$（$0 \leqq f(wb_{dq}) \leqq 1$）は，式 (2) によって算出できる．

$$f(wb_{dq}) = \frac{a \cdot x + \big(n_{BH}(wb_{dq}) + n_{BS}(wb_{dq})\big) \cdot p(wb_{dq})}{a + (n_{BH}(wb_{dq}) + n_{BS}(wb_{dq}))} \tag{2}$$

式 (2) において，x は新出の場合における共起語 wb_{dq} の有害度の初期値，a は初期値 x に与える強さを表す．a を用いることで，登場回数が少ない場合を考慮して wb_{dq} の有害確率を算出できる．

4. 判定アルゴリズム

4.1 Web ページの有害度の算出

本機能では，事前処理で構築した共起語辞書を用いて有害情報の判定を行う．Web ページの有害判定には，式 (3)〜(5) の Gary Robinson-Fisher 方式で提案されている有害判定指標 I_d を共起語の組合せに対応させた式 (6) の算出結果 I_{weight} を用いる．

$$H_d = C^{-1}\Big(-2\ln\prod\nolimits_{wt_{dg}}(1 - f(wt_{dg})), 2n\Big) \tag{3}$$

$$S_d = C^{-1}\Big(-2\ln\prod\nolimits_{wt_{dg}} f(wt_{dg}), 2n\Big) \tag{4}$$

$$I_d = \frac{1 + H_d - S_d}{2} \tag{5}$$

$$I_{weight} = \frac{1 \cdot I_1 + 2 \cdot I_2 + \cdots + d \cdot I_d}{1 + 2 + \cdots + d} \tag{6}$$

式 (3)，(4) において，C^{-1} は逆 χ^2 関数を表し，n は判定対象の Web ページ wt に出現する全語句数を表す．ここで，全語句数とは，$d = 1$ の場合は，判定対象に存在する全単語数，$d = 2$ の場合は，判定対象に存在する 2 つ組みの共起語数を指す．式 (3)，(4) における wt_{dg} は，判定対象から抽出した 1 語から d 語までの単語の組合せを表す．また，式 (3)〜(5) において，H_d は投稿の無害度，S_d は投稿の有害度，I_d（$0 \leqq I_d \leqq 1$）は S_d と H_d を統合して算出した Web ページの有害度指標をそれぞれ表す．I_d は，0 に近いほど判定対象の Web ページは無害である可能性が高く，1 に近いほど判定対象の Web ページは有害である可能性が高いと判断するための指標である．式 (6) において，d は共起の組合せ数を表し，I_d は式 (5) により算出した Web ページの有害度判定指標を表す．従来の Gary Robinson-Fisher 方式を用いる有害判定では，有害度判定指標 I_d の値を用いて判定対象の Web ページが有害か無害かを判定する．それに対して，本提案手法では，共起語の組合せ数を考慮するため，式 (6) に示す加重平均法で I_d を統合した結果である I_{weight} を用いて判定する．ここで加重平均法を用いた理由として，「会員 ログイン」の 2 つ組みの共起語よりも，「会員 ログイン 出会い 無料」など 4 つ組みの共起語の方が，判定対象の Web ページを特徴付ける要素であると考えたためである．なお，本提案手法では，I_{weight} が閾値 tv より大きい場合，判定対象の Web ページを有害情報と判定し，I_{weight} が閾値 tv 以下の場合，判定対象の Web ページを無害情報と判定する．

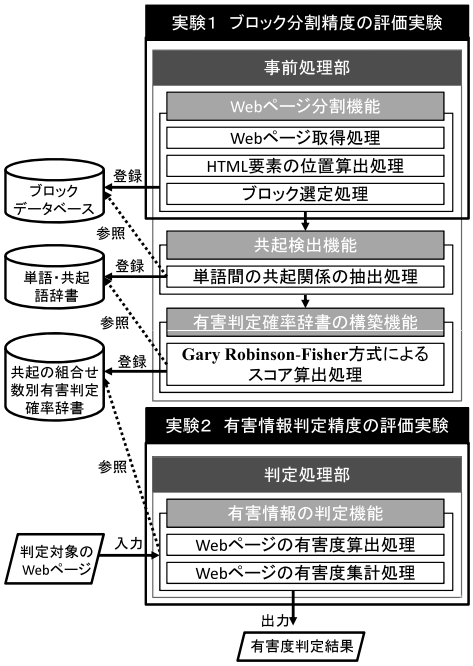

図 5　実験計画
Fig. 5　Plan of experiment.

5. 評価実験

5.1 実験計画

本研究で提案した有害情報判定手法の有用性を検証するため，「ブロック分割精度」と「有害情報判定精度」の 2 項目について，評価実験を行う．本研究の実験計画を**図 5** に示す．図 5 は，評価実験により検証する項目を明確化するため，図 1 と実験内容の対応関係を図に示したものである．実験 1 では，多様なウィンドウサイズに対応可能であるかを検証するため，複数の掲示板を対象としてブロック単位に分割し，その結果について考察する．実験 2 では，多様なウィンドウサイズを考慮した有害情報判定手法の有効性を証明するため，既存のウィンドウサイズを一定にした手法との比較実験を行い，その結果について考察する．

5.2 実験 1：ブロック分割精度の評価実験

5.2.1 実験内容

本実験では，ブロック分割の精度を評価するため，指定したドメインの Web ページに対して分割処理を適用し，ブロック単位に正しく分割されているかを検証する．実験に用いるデータは，Google [22] 検索の掲示板検索機能を用い

© 2013 Information Processing Society of Japan

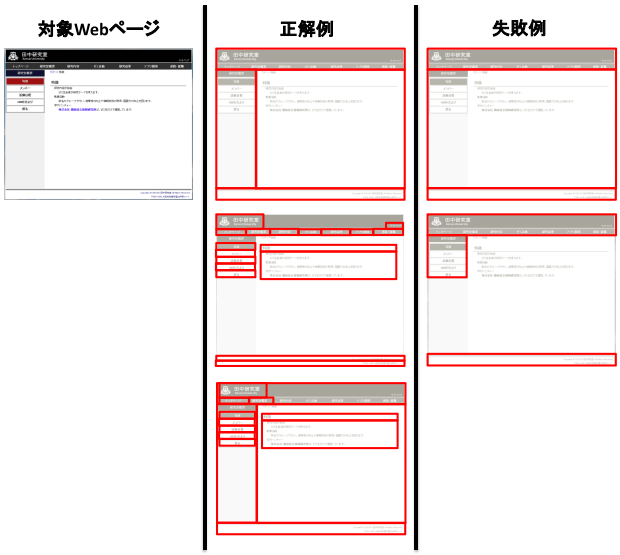

図 6　Web ページの分割結果の評価例
Fig. 6　Example of evaluation concerning segmentation of Web page.

て，出会いに関するキーワードで検索した結果，得られた Web ページとした．ここで，実験対象として電子掲示板を採用した理由は，電子掲示板がユーザ参加型のコンテンツであり，有害情報に関する投稿が多く見られるためである．また，同様のドメインの Web ページは，同一の HTML 構造であることから，ドメイン単位で Web ページの分割が可能であったかを評価する．Web ページ分割の正否判定では，Web ページの分割結果を目視で確認し，ページがブロック単位に分割されていれば正解とする．Web ページの分割結果についての正否の判定例を**図 6** に示す．本実験では，図に示すとおり，Web ページをメニューやヘッダ，メインコンテンツなどを漏れなくブロック単位に分割できている場合に正解と判断する．本実験の手順を次に示す．

STEP 1.　Google 検索の掲示板検索機能を用いて，援助交際目的の出会いに頻繁に利用される 4 つのキーワード（苺佐保，ホ別，援助交際，W 諭吉）で検索し，各キーワードに一致した Web ページを取得する．

STEP 2.　収集した Web ページをドメイン単位にグループ化する．これは，実験対象として，同一のドメインの Web サイトであれば，同様の HTML 構造となっており，同一の判定結果が得られると考えたためである．本ステップでは，ドメイン数が 150 件になるまで繰り返し実施する．

STEP 3.　ドメイン単位にグループ化した Web ページのレイアウト例（**図 7**）に基づき分類する．

STEP 4.　レイアウトの構図ごとにグループ分けしたドメインから，実験対象となる Web ページを取得し，ブロック単位に分割する．

STEP 5.　分割結果を目視で確認し，Web ページ分割の精度を算出する．

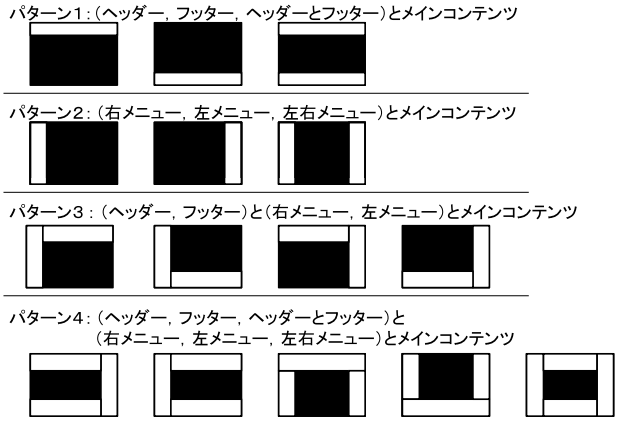

図 7　Web ページのレイアウト例
Fig. 7　Example of layout of Web page.

表 1　「最大の要素面積」とヘッダ，フッタ，メニューの面積の割合
Table 1　Results compared *MaxElementArea* with area of header, footer and menu.

	ヘッダー	フッター	メニュー
最小	1.55%	0.29%	7.01%
最大	9.67%	9.25%	18.54%
面積の割合（平均）	5.14%	4.27%	14.13%

5.2.2　ブロック分割精度の評価実験用パラメータの設定

ブロック分割精度の評価実験では，Web ページ分割アルゴリズムにおける閾値であるパラメータ α，β を用いる．各パラメータについて，次に示すとおり設定した．

(1) パラメータ α

パラメータ α は，ページ全体の大部分を占めるタグが複数定義されており，さらに内部のコンテンツの量が少ない場合に，ページ全体を 1 つのブロックとして抽出される現象を抑制するための閾値である．本パラメータは，ヘッダ，フッタ，メニューやメインコンテンツなどをすべて含む大枠のタグを除去可能なように値を設定する必要がある．ページ全体の大部分を占める大枠のタグは，ヘッダ，フッタ，メニューとメインコンテンツなどの Web ページを構成する要素を含むサイズからメインコンテンツのサイズまでの間のサイズである．そのため，Web ページ全体のサイズからヘッダ，フッタもしくはメニューのサイズを差し引いた値を設定すれば，大枠のタグを除去できると考えられる．本研究では，Web ページを構成するヘッダ，フッタもしくはメニューのページに占める割合を調査し，調査結果に基づきパラメータ α の値を設定する．Web ページ 50 件を対象として，Web ページ内の最大のタグの面積に対するヘッダ，フッタ，メニューの面積割合を調査した結果を**表 1** に示す．調査結果を確認すると，ヘッダ，フッタは最大 10%未満，メニューは 20%未満であることが分かる．また，図 7 のパターンに示すとおり，主要な Web ページ

© 2013 Information Processing Society of Japan

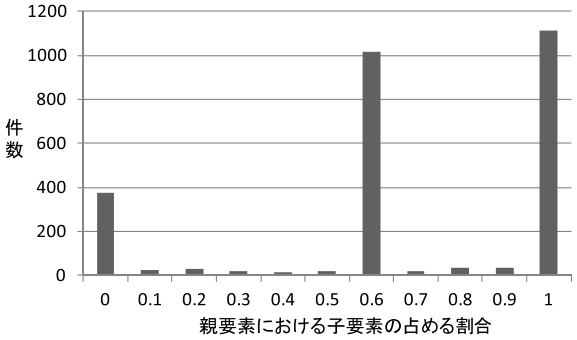

図 8　「親要素」における「子要素」の占める割合の分布
Fig. 8　Distribution of *ChildElements* into *ParentElement*.

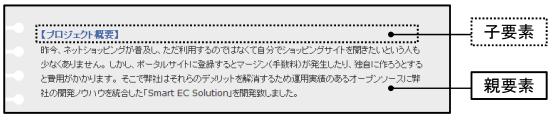

図 9　0.4 未満の「子要素」の領域例
Fig. 9　Area of *ChildElements* less than 0.4.

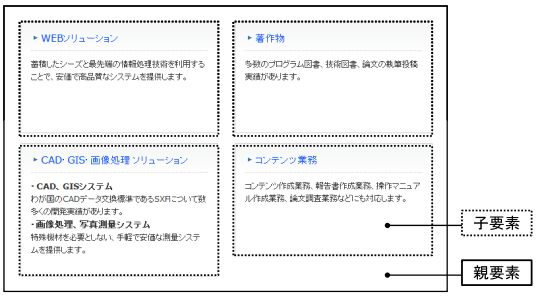

図 10　0.6 以上の「子要素」の領域例
Fig. 10　Area of *ChildElements* over 0.6.

のレイアウトにはヘッダもしくはフッタが含まれていることが分かる．そのため，大枠のタグをヘッダ（フッタ）とメインコンテンツを合算したサイズより小さく，メインコンテンツのサイズより大きいものと考え，本実験ではパラメータ $\alpha = 0.9$ と設定する．

(2) パラメータ β

パラメータ β は，子要素の占める割合が自身の領域の多くを占める場合，子要素それぞれを独立したブロック要素として抽出するための閾値である．本研究では，パラメータ β の値を適切に設定するため，2,711 件の親 HTML 要素における子 HTML 要素の占める割合の分布を調査（**図 8**）した．調査結果を確認すると，0.4 未満が約 17%，0.6 以上が約 82%であり合算して全体の約 99%を占めており，0.4 以上 0.6 未満の範囲は約 1%と一部の要素のみになっていることが分かる．また，0.4 未満のレイアウトの例（**図 9**）や 0.6 以上のレイアウトの例（**図 10**）を確認すると，親領域の 0.4 未満もしくは 0.6 以上の場合においては，領域が大きい方が主要なコンテンツであることが分かる．そのため，本研究では，親領域の半数を超える場合は子領域をブロックとして抽出可能とするため，パラメータ $\beta = 0.5$ と設定する．

表 2　Web ページの分割精度
Table 2　Accuracy of segmentation of Web page.

レイアウトの構図	データ件数	正解件数	正解率
パターン 1	63	54	0.8571
パターン 2	1	1	1.0000
パターン 3	8	5	0.6250
パターン 4	78	76	0.9744
全体	150	136	0.9067

表 3　誤判定したデータの分析
Table 3　Analysis of Incorrect data.

分割ミスの原因		パターン 1	2	3	4
HTML 要素取得時の問題	<hr>タグ， タグによるレイアウト	2	0	0	0
	<font>タグによるレイアウト	2	0	0	0
	<body>のみ	3	0	0	0
レイアウト判定ミス		0	0	3	1
処理エラー		2	0	0	1

5.2.3　実験結果と考察

Web ページ分割実験の結果を**表 2** に示す．電子掲示板のドメイン 150 件を分類した結果，パターン 1 が 63 件（約 42%），パターン 4 が 78 件（約 52%）となり，全体の約 9 割がパターン 1 か 4 に分類されることが分かった．また，それぞれの Web ページの分割精度を確認するとそれぞれ 0.8571（パターン 1），0.9744（パターン 4）であり，約 9 割のドメインは正しく分割できることが明らかになった．

Web ページの分割が失敗した 14 ドメインを失敗原因に基づき分類した結果を**表 3** に示す．分類結果を確認すると半数のドメインが HTML 要素を正しく認識できず，分割に失敗していることが明らかになった．具体的には，
 タグと <hr> タグ，もしくは <font> タグを用いてレイアウトを作成している Web ページや，<body> タグ直下にレイアウト要素が存在しない Web ページが見られた．この課題に対して，HTML 要素間の包含関係に着目する本提案手法では対応できない状況である．これらの Web ページに対しては，HTML の繰返し構造に着目して出現パターンを学習し，その結果に基づき分割することで対応できると考えられる．

5.3　実験 2：有害情報判定精度の評価実験

5.3.1　実験内容

本実験では，本研究で提案した多様なウィンドウサイズを考慮した有害情報判定手法の有効性を検証するため，あらかじめ人手で取得した Web ページ 4,000 件（出会い系サ

© 2013 Information Processing Society of Japan

イトを含む有害ページ 2,000 件，無害ページ 2,000 件）を判別し，その精度を評価する．判定精度の評価実験では，多様なウィンドウサイズの有効性を検証するため，次に示す 3 つの手法を用いる．

- Web ページ全体から単語を抽出する．そして，抽出した単語の有害度のスコアを Gary Robinson-Fisher 方式により算出し，その結果を用いて有害情報を判定する手法（以下，「単語判定手法」）．
- Web ページ全体を共起関係抽出のウィンドウサイズとして，単語と共起語を抽出する．そして，抽出した単語と共起語の有害度のスコアを Gary Robinson-Fisher 方式により算出し，その結果を用いて有害情報を判定する手法（以下，「ウィンドウサイズが一定の判定手法」）．
- Web ページをブロック単位に分割した結果をウィンドウサイズとして，単語と共起語を抽出する．そして，抽出した単語と共起語の有害度のスコアを Gary Robinson-Fisher 方式により算出し，その結果を用いて有害情報を判定する手法（以下，「ブロック分割手法」）．

本実験の手順を次に示す．

STEP 1. 有害ページ 2,000 件，無害ページ 2,000 件の文章を MeCab により形態素解析し，名詞，形容詞，動詞を単語として抽出する．

STEP 2. 各 Web ページに共通して出現する単語の組合せを抽出する．

STEP 3. Web ページ 4,000 件を 5 分割し，有害ページ 400 件，無害ページ 400 件となる 5 つのデータセットを作成する．

STEP 4. 5 つのデータセットの 1 つを判別対象データ，残りの 4 つを教師データとする．

STEP 5. 教師データの Web ページをブロック単位に分割し，分割結果に含まれる単語の組合せのみに絞り込みを行う．

STEP 6. 教師データの Web ページの STEP 1 の結果に対して，単語判定手法を適用し，有害語と無害語の辞書を構築する．

STEP 7. 教師データの Web ページの STEP 1，STEP 2 の結果に対して，ウィンドウサイズが一定の判定手法を適用し，有害語と無害語の辞書を構築する．

STEP 8. 教師データの Web ページの STEP 1，STEP 5 の結果に対して，ブロック分割手法を適用し，有害語と無害語の辞書を構築する．

STEP 9. STEP 6〜STEP 8 で構築した 3 つの辞書を用いて，判別対象データのデータセット（有害ページ 400 件，無害ページ 400 件）を判定し，その精度を評価する．なお，判別精度の評価指標には，情報検索の精度評価に一般的に用いられる F 値を用いる．

STEP 10. STEP 4 で選択した判別対象データと異なるデータセットを判定対象データとして，STEP 4〜STEP 9 を繰り返し実施する（5 分割交差法の実施）．

5.3.2 有害情報判定精度の評価実験用パラメータの設定

有害情報判定精度の評価実験では，共起語の組合せ数を表現するパラメータ d，Gary Robinson-Fisher 方式で用いるパラメータ x，a と有害判定指標 I_{weight} の閾値 tv を用いる．各パラメータについて，次に示すとおり設定した．

(1) パラメータ d

パラメータ d は，ウィンドウサイズが一定の判定手法とブロック分割手法において，共起語の組合せ数を設定するために用いる．本研究では，パラメータ d の値を設定するために，d の値を変化させて精度を評価し，有害情報の判定に適した値を決定する．本実験では，評価実験で用いる実験データを対象として，ウィンドウサイズが一定の判定手法で有害情報判定した際の F 値をもとにパラメータを決定する．なお，共起語の組合せ数 d の値は，1，2，3，4 の 4 種類とする．

パラメータの決定実験の結果を**図 11** に示す．実験結果を確認すると，共起語の組合せが 2 つ組の場合が最も精度が良く，3 つ組，4 つ組と増加するに従って，判定精度が低下していることが分かる．詳細を確認すると，共起語の組合せ数が増加するに従って，有害ページを無害ページとして判定する数が増加していることが明らかになった．これは，Web ページ内に登場する単語の組合せと，辞書に登録されている有害語の組合せが一致せず，有害判定指標の値が低下したためであると考えられる．そのため，パラメータの決定実験の結果から，本実験では最も精度の良い $d = 2$ を用いて判定精度を算出する．

(2) パラメータ x と a

パラメータ x，a は，Gary Robinson-Fisher 方式で文書に出現した語句の有害確率を算出する際に用いる．パラメータ x は，文書に出現した語句が新出の場合における有害度の初期値，パラメータ a は，パラメータ x に与える強さを表す．これらのパラメータは，Gary Robinson-Fisher 方式を提案した文献 [20] において，$x = 0.5$，$a = 1$ を与えた場合に有効な結果が得られると述べている．そのため，

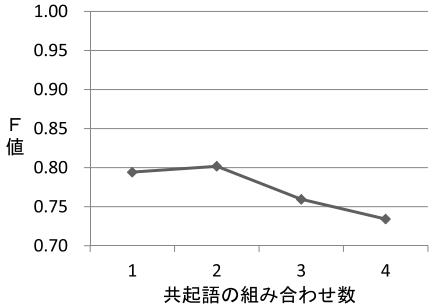

図 11 F 値とパラメータ d との関係
Fig. 11 Relationship between F-measure and parameter d.

© 2013 Information Processing Society of Japan

表 4 各手法の有害情報の判定精度
Table 4 Accuracy of harmful information by each method.

手法		無害ページ	有害ページ	全体
単語判定手法	適合率	0.7775	0.9959	0.8503
	再現率	0.9970	0.7035	0.8503
	F 値	0.8719	0.8178	0.8503
ウィンドウサイズが一定の判定手法	適合率	0.7253	0.9992	0.8035
	再現率	0.9995	0.6075	0.8035
	F 値	0.8386	0.7454	0.8035
ブロック分割手法	適合率	0.9306	0.9910	0.9575
	再現率	0.9915	0.9235	0.9575
	F 値	0.9595	0.9552	0.9575

本研究においても同様に，$x = 0.5$，$a = 1$ として設定する．

(3) パラメータ tv

パラメータ tv は，有害情報判定指標 I_{weight} の値に基づき対象 Web ページが有害か無害であるかを判定する際の閾値である．有害情報判定指標 I_{weight} の値がパラメータ tv よりも大きい場合は有害ページ，tv 以下の場合は無害ページであると判定する．パラメータ tv は，Gary Robinson-Fisher 方式を用いた同様の取り組みの研究 [8], [12] に倣い，$tv = 0.5$ と設定する．

5.3.3 実験結果と考察

有害情報の実験用データセットを対象として，単語判定手法，ウィンドウサイズが一定の判定手法とブロック分割手法のそれぞれの手法で判定した結果を**表 4** に示す．実験結果を確認すると次に示す 3 つの内容が明らかになった．

これら 3 つの特徴から，有害情報の判定手法として，多様なウィンドウサイズに対応した提案手法であるブロック分割手法は，既存手法である単語判定手法やウィンドウサイズが一定の判定手法と比較して，高精度に有害情報を判定できることが明らかになった．

• ブロック分割手法が最も高精度に有害情報を判定可能

実験結果（表 4）を確認すると，単語判定手法の F 値が 0.8503，ウィンドウサイズが一定の判定手法の F 値が 0.8035，ブロック分割手法の F 値が 0.9575 となっており，ブロック分割手法が最も高精度に有害情報を判定できることが明らかになった．

まず，ブロック分割手法とウィンドウサイズが一定の判定手法とを F 値で比較すると，0.1540 ポイントの差で，ブロック分割手法が高精度に判定できていることが分かる．このことから，ウィンドウサイズを多様にした方が，高精度に有害情報を判定できることが明らかになった．

次に，ブロック分割手法と単語判定手法とを F 値で比較すると，0.1072 ポイントの差で，ブロック分割手法が高精度に判定できていることが分かる．しかし，ウィンドウサイズが一定の判定手法と単語判定手法とを F 値で比較すると，0.0468 ポイントの差で単語判定手法が高精度に判定できていることが分かる．このことから，ウィンドウサイズが一定の判定手法では単語の共起関係を考慮することの効果は得られず，ウィンドウサイズを多様にしたうえで，単語の共起関係を考慮することが重要であることが明らかになった．ブロック分割手法とウィンドウサイズが一定の判定手法の判定精度差の要因を分析するため，各手法の有害語辞書を比較した結果，ブロック分割手法のインデックス数 128,267 件，ウィンドウサイズが一定の判定手法のインデックス数 158,674 件となっており，ウィンドウサイズが一定の判定手法の方が，30,407 件多いことが分かった．各辞書の有害度上位 1,000 件を比較してウィンドウサイズが一定の判定手法の辞書にのみ存在した共起語の例を**表 5** に示す．共起語の例を確認すると地名の組合せや Web ページに共通して存在する語句（サイトマップの「マップ」や jpg，戻る，コチラなど）との組合せが抽出されており，誤判定につながる可能性の高い共起語が辞書に含まれていることが分かる．

表 5 「ウィンドウサイズが一定の判定手法」による有害語辞書に存在した共起語の例
Table 5 Example of word co-occurrence existed in harmful word dictionary using detection method of fixed window size.

キャリア　コチラ	期間お　上尾
A　コミュニティー	下関　jpg
消えろ　戻る	マップ　Mobile
ある　鯖江	戻る　上尾
新横浜　at	Mobile　SEARCH
キャッシング　待ち合わせ	下関　上尾
家政　通りすがり	友人　知人
市内　通りすがり	大阪　奈良
兵庫　京都	稲城　伊豆
京都　奈良	上尾　歩い
羽生　上尾	小平　マップ
下関　小平	兵庫　大阪
上尾　長浜	上尾　伊豆
大野　上尾	小平　評判
稲城　古川	
小平　大野	

これらの結果をまとめると，単語の共起関係を考慮するのみでは精度が向上せず，多様なウィンドウサイズを考慮する手法と組み合わせることで，大幅な精度向上が見られることが明らかになった．これは，各手法の有害語辞書を分析した結果から，多様なウィンドウサイズを考慮することで共起を抽出する範囲が限定され，誤判定につながる共起語の抽出が抑制されたためであることが明らかになった．

© 2013 Information Processing Society of Japan

表 6 ブロック分割手法とウィンドウサイズが一定の手法の有害語辞書上位 1,000 件の一致数
Table 6 Number of matches of top 1,000 of harmful word dictionary between block segmentation method of proposal method and detection method of fixed window size.

	100 位以内	500 位以内	1000 位以内
データセット 1	98	474	752
データセット 2	100	469	661
データセット 3	97	352	473
データセット 4	100	479	718
データセット 5	100	462	603
平均	99.00	447.20	641.40

- **ブロック分割手法は有害ページを無害ページと判定した割合が最も低い**

実験結果（表 4）を確認すると，単語の判定手法およびウィンドウサイズが一定の判定手法では，ブロック分割手法と比較して無害ページの適合率と有害ページの再現率が低い値となっている．無害ページの適合率が高い場合，無害と判定した Web ページの中に，有害の Web ページが含まれる割合が少ないことを表している．また，有害ページの再現率が高いと有害ページを正しく有害と判定しており，有害ページの判定漏れが少ないことを表している．このことから，単語の判定手法およびウィンドウサイズが一定の判定手法は，有害ページを無害ページと誤判定する可能性が高い状況である．

ブロック分割手法が高精度に有害ページを判定可能な要因を解析するため，各手法の有害語辞書を比較した．具体的には，ブロック分割手法の 2 つ組み有害語辞書の有害度上位 1,000 件に対して，ウィンドウサイズが一定の判定手法の有害語辞書の有害度上位 1,000 件が何件一致するかを検証した．結果（**表 6**）を確認すると，上位 1,000 件中平均すると 641.40 件（約 64%）となっており，各手法の共起語の有害語辞書が異なっていることが明らかになった．ブロック分割手法の辞書の有害度上位 1,000 件にのみ存在した共起語の一例を**表 7** に示す．結果を確認すると，「大阪　困っ」や「ください　助け」など，一見して無害な共起語と判断される可能性の高い語句の有害度が高くなっていることが分かった．しかし，表 7 に示す実際の書き込み記事の例を確認すると，男性が女性を誘う有害な記事などにおいても頻繁に利用されていることが分かる．このことから，ブロック分割手法では，これらの共起語とそのページに存在する他の有害度の高い語（出会い，援助，神待ちなど）を用いて，Web ページの有害度判定指標を算出しているため，有害な Web ページを正しく判定できていると考えられる．

表 7 ブロック分割手法による有害語辞書に存在した共起語の例
Table 7 Example of word co-occurrence existed in harmful word dictionary using block segmentation method (proposal method).

共起語	有害度	書き込み記事の例
大阪　困っ	0.9854	大阪で泊まる場所困ってる子宿泊できますよ
女子　場所	0.9780	女子高生　└食事・泊まる場所提供出来る男性を募集
女性　困っ	0.9759	家出とかで泊まるところがなく本当に困っている女性だけ mail 下さい
ください 女の子	0.9810	かまってほしい女の子いたら、メールください
ください 助け	0.9759	助けて欲しい人メールください

- **共起語を用いることで有害判定指標 I_{weight} の値の分布に偏りが見られる**

各手法で実験データの Web ページを判定した際の有害判定指標 I_{weight} の分布を**表 8** に示す．有害判定指標 I_{weight} の分布を確認すると，単語判定手法では，有害ページのスコアが分散しており，0.10 から 0.90 まで幅広く分布していることが分かる．それに対して，Web ページの判定に共起語を用いるウィンドウサイズが一定の判定手法およびブロック分割手法は，0.15 以下もしくは，0.85 以上に偏っていることが分かる．これは，共起関係を考慮することで，Web ページに出現する単語の特徴をより強調できたためであると考えられる．

6. おわりに

本研究では，有害情報フィルタリングのための特徴抽出の処理範囲であるウィンドウサイズが一定であるという問題を解消するため，Web ページの見た目の特徴に基づきブロック単位に分割し，そのブロックをウィンドウサイズとする手法を提案した．本提案手法についての有用性を評価するため，ページ全体を共起語抽出のウィンドウサイズとした場合との比較実験を実施した結果，F 値で 0.1540 ポイントの差で本提案手法の方が高精度に有害情報を判定できることが明らかになった．この結果から，ウィンドウサイズをブロック単位とした本提案手法の有害情報フィルタリングの有用性を実証した．

今後は，実証実験で明らかになった Web ページをブロック単位に分割する際の課題を解消した新たな Web ページ分割手法を提案する予定である．具体的には，Web ページからブロックを抽出する段階において，ユーザ操作で動的に HTML 要素を作成して DOM を更新する Web ページや，<hr> タグや
 タグのみでレイアウトが構成され

© 2013 Information Processing Society of Japan

表 8　各手法の有害判定指標 I_{weight} の度数分布
Table 8　Frequency distribution of score I_{weight} by each method.

スコア	単語判定手法		ウィンドウサイズが一定の判定手法		ブロック分割手法	
	無害	有害	無害	有害	無害	有害
0.00	774	16	595	428	606	92
0.05	1,062	275	1,349	278	1,299	37
0.10	1	33	40	29	39	7
0.15	0	30	15	26	36	5
0.20	0	24	0	24	1	2
0.25	2	23	0	0	0	0
0.30	0	35	0	0	1	0
0.35	0	21	0	0	0	8
0.40	0	33	0	0	0	0
0.45	0	23	0	0	1	2
0.50	155	80	0	0	0	0
0.55	3	33	0	0	0	0
0.60	0	31	0	0	0	0
0.65	0	27	0	0	0	0
0.70	2	52	0	7	2	3
0.75	0	45	0	0	0	0
0.80	0	41	0	0	0	0
0.85	0	61	0	61	2	9
0.90	0	78	0	85	0	39
0.95	0	107	0	132	2	45
1.00	1	932	1	930	11	1,751

ている Web ページに対応した分割手法を検討する．また，本研究の応用として，Web ページの分割手法を有害情報の判定処理時に適用する方法を検討する予定である．このことにより，Web ページに部分的に有害情報が含まれている場合（たとえば，有料広告の一部のコンテンツなど）に，Web ページの一部分のみを有害情報としてフィルタリングできるため，Web ページ全体が有害情報と判断されて閲覧できないという課題を解消できると考えられる．

謝辞　本研究の一部は，平成 21～22 年度 JST RISTEX「犯罪からの子どもの安全」研究開発領域研究開発プログラム「犯罪からの子どもの安全」（研究課題「子どもの犯罪に関わる電子掲示板記事の収集・監視手法の検討」），平成 20～24 年度私立大学戦略的研究基盤形成支援事業（研究課題「セキュアライフ創出のための安全知循環ネットワークに関する研究」）から助成を受け，その成果を公表するものである．

参考文献

[1] 文部科学省：子どもの携帯電話等の利用に関する調査 (2009).
[2] 内閣府：青少年が安全に安心してインターネットを利用できる環境の整備等に関する法律 (2008).
[3] Lee, W., Lee, S.S., Chung, S. and An, D.: Harmful Contents Classification Using the Harmful Word Filtering and SVM, *Proc. 7th International Conference on Computational Science*, pp.18–25, Springer-Verlag (2007).
[4] Guermazi, R., Hammami, M. and Hamadou, A.: Combining Classifiers for Web Violent Content Detection and Filtering, *Proc. 7th International Conference on Computational Science*, pp.773–780, Springer-Verlag (2007).
[5] Lee, P.Y., Hui, S.C. and Fong, A.C.M.: Neural Networks for Web Content Filtering, *IEEE Intelligent Systems*, Vol.17, No.5, pp.48–57 (2002).

© 2013 Information Processing Society of Japan

[6] Du, R., Safavi-Naini, R. and Susilo, W.: Web Filtering Using Text Classification, *IEEE International Conference on Networks*, Vol.11, pp.325–330 (2003).

[7] Chandrinos, K.V., Androutsopoulos, I., Paliouras, G. and Spyropoulos, C.D.: Automatic Web Rating: Filtering Obscene Content on the Web, *Proc. 4th European Conference on Research and Advanced Technology for Digital Libraries*, pp.403–406, Springer-Verlag (2000).

[8] 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講演論文集，No.2, pp.1–6, 情報処理学会・電子情報通信学会 (2010).

[9] 池田和史，柳原　正，松本一則，滝嶋康弘：HTML 要素に着目した違法・有害サイト検出手法の提案と評価，第 9 回情報科学技術フォーラム講演論文集，Vol.FIT2010, No.2, pp.7–12, 情報処理学会・電子情報通信学会 (2010).

[10] 松葉達明，里見尚宏，桝井文人，河合敦夫，井須尚紀：学校非公式サイトにおける有害情報検出，言語理解とコミュニケーション研究会技術研究報告，Vol.109, No.142, pp.93–98, 電子情報通信学会 (2009).

[11] 中村健二，田中成典，大谷和史，山本雄平：セキュアライフの創出を目指した安全知の獲得に関する研究—電子掲示板からの犯行予告の抽出，土木情報利用技術論文集，Vol.18, pp.269–280, 土木学会 (2009).

[12] 吉村卓也，藤井雄太郎，伊藤孝行：Robinson 型判定手法を用いた単語共起フィルタの検証，第 10 回情報科学技術フォーラム講演論文集，No.2, pp.85–90, 情報処理学会・電子情報通信学会 (2011).

[13] 石坂達也，山本和英：2 ちゃんねるを対象とした悪口表現の抽出，第 16 回年次大会発表論文集，pp.178–181, 言語処理学会 (2010).

[14] 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，*DEIM Forum 2010*, 日本データベース学会 (2010).

[15] Cortes, C. and Vapnik, V.: Support-Vector Networks, *Machine Learning*, Vol.20, No.3, pp.273–297, Springer-Verlag (1995).

[16] Duda, R.O. and Hart, P.E.: *Pattern Classification and Scene Analysis*, John Willey & Sons (1973).

[17] Wiener, E., Pedersen, J.O. and Weigend, A.S.: A Neural Network Approach to Topic Spotting, *Proc. 4th Annual Symposium on Document Analysis and Information Retrieval, SDAIR*, pp.317–332 (1995).

[18] Doug, B., Adam, B. and John, L.: Statistical Models for Text Segmentation, *Machine Learning*, Vol.34, pp.177–210, Springer-Verlag (1999).

[19] Burget, R.: Web Page Element Classification Based on Visual Features, *Proc. 1st Asian Conference on Intelligent Information and Database Systems 2009*, pp.67–72, IEEE (2009).

[20] Robinson, G.: A Statistical Approach to the Spam Problem, Specialized Systems Consultants, *Linux Journal*, Vol.107, pp.58–64 (2003).

[21] 工藤　拓：MeCab，入手先 ⟨http://mecab.sourceforge.jp/⟩（参照 2012-05-14）.

[22] Google Inc.: Google, available from ⟨http://www.google.co.jp/⟩（参照 2012-05-14）.

中村 健二　（正会員）

1981 年生．2004 年関西大学総合情報学部卒業．2006 年関西大学大学院総合情報学研究科知識情報学専攻博士課程前期課程修了．2009 年関西大学大学院総合情報学研究科総合情報学専攻博士課程後期課程修了．同年関西大学ポスト・ドクトラル・フェロー，2010 年立命館大学情報理工学部助手，2012 年大阪経済大学情報社会学部准教授　現在に至る．博士（情報学）．知識情報処理，Web マイニング，テキストマイニング等の研究に従事．2002 年から（株）関西総合情報研究所で活動．システム設計，データモデル設計等の研究開発に従事．電子情報通信学会，土木学会，日本データベース学会各会員．

田中 成典　（正会員）

1963 年生．1986 年関西大学工学部土木工学科卒業．1988 年関西大学大学院工学研究科土木工学専攻博士課程前期課程修了．同年（株）東洋情報システム（現在，TIS）に入社．人工知能に関する研究受託開発業務に従事．1994 年関西大学総合情報学部専任講師．1997 年助教授，2004 年教授，2006 年から学生センター副所長，現在に至る．2002 年 8 月から 1 年間，カナダの UBC で客員助教授．博士（工学）．専門は知識工学と社会基盤情報学．CAD/CG，GIS/GPS，画像処理および Web ソリューションビジネスに関する研究に従事．2000 年（株）関西総合情報研究所を起業，設立当初から現在まで取締役会長．2006～2012 年（株）フォーラムエイトの顧問．建設省土木研究所 CAD 製図基準検討委員会委員長，土木学会土木情報システム委員会幹事長，同委員会土木 CAD 小委員会委員長，ISO/TC184/SC4 国内委員等を歴任．現在，国土交通省日本建設情報総合センター社会基盤情報標準化委員会委員，同委員会 CAD/データ連携小委員会委員長，土木学会情報利用技術委員会副委員長．主に，ISO に準拠した CAD 製図基準と CAD データ交換基盤の開発に従事．

© 2013 Information Processing Society of Japan

山本 雄平　（学生会員）

1986 年生．2009 年関西大学総合情報学部総合情報学科卒業．2011 年関西大学大学院総合情報学研究科知識情報学専攻博士課程前期課程修了．現在，関西大学大学院総合情報学研究科総合情報学専攻博士課程後期課程在学中．修士（情報学）．Web マイニング，自然言語処理の研究に従事．2007 年（株）関西総合情報研究所入社．現在に至る．システム設計等の研究開発に従事．

安彦 智史　（学生会員）

1986 年生．2008 年関西大学総合情報学部総合情報学科卒業．2010 年関西大学大学院総合情報学研究科知識情報学専攻博士課程前期課程修了．現在，関西大学大学院総合情報学研究科総合情報学専攻博士課程後期課程在学中．修士（情報学）．画像処理，自然言語処理の研究に従事．2006 年（株）関西総合情報研究所入社．現在に至る．システム設計等の研究開発に従事．

© 2013 Information Processing Society of Japan