情報処理学会論文誌 Vol. 52 No. 8 2474–2483 (Aug. 2011)

推薦論文

HTML要素に基づく有害サイト検出手法

池田和史[†1] 柳原正[†1] 服部元[†1]
松本一則[†1] 小野智弘[†1] 滝嶋康弘[†1]

本稿では高速かつ高精度に有害サイトを検出するため，Webサイトの背景色やリンク先，ブラウザに特定の動作をさせるスクリプトなど，有害サイトに特徴的に見られる傾向をHTML要素から検出する手法を提案する．提案手法では有害サイトのHTMLに偏って出現するような文字列を自動的に抽出し，SVM（Support Vector Machine）を用いてこれらの特徴を組み合わせて有害サイトの検出を行う．提案手法はWebサイトの本文の情報を利用しないため，既存のキーワードベース方式によって検出が困難なサイトも検出が可能である．このため，既存のキーワードベース方式と組み合わせて利用することで検出精度を向上させることも可能である．大規模なWebサイトデータを用いた性能評価実験を行い，既存のキーワードベース方式と比較して，適合率を9.3ポイント向上するなどの性能向上を確認した．

Detection of Malicious Web Pages Based on HTML Elements

Kazushi Ikeda,[†1] Tadashi Yanagihara,[†1]
Gen Hattori,[†1] Kazunori Matsumoto,[†1]
Chihiro Ono[†1] and Yasuhiro Takishima[†1]

In this paper, we propose high-speed and accurate algorithms for detecting malicious Web pages. Our algorithms detect the features of malicious Web pages from their HTML elements such as the background colors of Web pages, the server names related to malicious Web pages, or the name of javascript functions that makes browsers perform unusual actions in response to malicious Web pages. Strings that appear especially in HTML elements of malicious Web pages are automatically chosen. SVMs (Support Vector Machines) combine these strings and detect malicious Web pages. Since our algorithms do not rely on the text parts of Web pages, they can detect Web pages that existing text-based algorithms have difficulty in detecting. By conducting a large-scale performance evaluation with real malicious Web pages, we showed that the hybrid algorithms of our algorithms and existing text-based algorithms increase the precision of existing text-based algorithms alone by 9.3 points.

1. まえがき

インターネットの普及により，一般ユーザ向けのWebサイトや掲示板が増加している．出会い系サイトや犯罪予告サイト，誹謗・中傷などの書き込みを含む学校裏サイトなど，有害な情報を含むサイトも増加傾向にあり，目視によるサイトの監視に要するコストは大きなものとなっている．近年，有害なWebサイトを自動的に検出するためのフィルタリングシステムの開発が進んでおり，ウェブブラウザに組み込まれてリアルタイムに有害サイトを検出する応用や，Webサイトの監視事業者が膨大なWebサイトの中から有害性の高いサイトを優先的に目視により監視する応用などが想定されるため，高精度かつ高速な判定が可能な有害サイト検出手法が求められる．

既存の主流な有害サイト検出手法としてWebサイトのURLを利用するBlack/Whiteリスト方式があるが，データベースを管理する人的コストが大きい点や，ブログなどでは同一ドメイン下に有害サイトと無害サイトの両方が存在する場合があるために判定精度が低下する点，新規のサイトに対して判定が行えない点などが課題としてあげられる．これに対し，Webサイトに記載の文書や掲載された画像を解析し，文書に特定のキーワードが含まれていることや画像の特徴を利用することで，有害サイトを検出するコンテンツベースの手法も提案されているが，単純な方式では高精度に有害サイトを検出することは難しく，一方で高度な言語処理や画像処理を行う手法では処理時間が大きくなることが課題である．

総務省が2008年に実施した調査[1)]によると，インターネット上で公開されている国内のブログの総数は1,690万ブログ（記事総数は13億5,000万記事）存在し，毎月4,000万記事が新規に投稿されている．有害な記事の割合はブログの運営事業者により異なるが，たとえば全体の10%が有害な記事であると仮定し，監視事業者が有害な400万記事のうち280万記事を発見，削除するというタスクを考える（再現率は70%となる）．フィルタリングシ

†1 株式会社KDDI研究所
KDDI R&D Laboratories Inc.
本稿の内容は2010年9月のFIT2010第9回情報科学技術フォーラムにて報告され，同プログラム委員長により情報処理学会論文誌ジャーナルへの掲載が推薦された論文である．

ステムの適合率は一般に 100%に満たないため，監視事業者は無害な記事を誤って有害と判定して削除しないように，最終的には人手で目視を行った後に記事を削除するが，このときフィルタリングシステムによって有害性が高いと判定された記事から優先的に目視を行うことで，作業を効率化することができるものと想定される．ここで，フィルタリングシステムの適合率が 60%の場合，有害な 280 万記事を発見するには 280 万/60% = 467 万記事を目視により確認する必要がある．このとき，無害な記事を 187 万記事確認することになるが，適合率が 70%の場合，280 万/70% = 400 万記事の確認により目標を達成でき，無害な記事は 120 万記事しか確認せずに済む．目視可能な記事数を 1 万記事/人日とすると，削減可能な人的コストは大きい．また，フィルタリングシステムにおける処理時間の短縮も運営設備の削減などコストの削減につながる．

本稿では高速かつ高精度に有害サイトを検出するため，Web サイトの HTML を対象とした有害サイト検出手法を提案する．提案手法では有害サイトの HTML に偏って出現するような文字列を自動的に抽出し，SVM（Support Vector Machine）を用いてこれらの特徴を組み合わせて有害サイトの検出を行う．提案手法は Web サイトの本文の情報を利用しないため，既存のキーワードベース方式によって検出が困難なサイトも検出が可能である点が特徴である．このため，既存のキーワードベース方式と組み合わせて利用することも有効である．

以降，2 章において提案手法に関連する研究について述べ，3 章において本稿における有害サイトの定義を記載し，4 章において提案手法の詳細について説明する．5 章では，既存のキーワードベース方式による有害サイトの検出手法と提案手法を組み合わせることによる有効性について検討し，6 章において提案手法に対する性能評価実験の内容と実験結果に対する考察を行う．

2. 関 連 研 究

Web サイトに記載の文書情報を利用して有害サイトを自動的に検出するいくつかの手法が提案されている[2),3)]．文献 2) の手法では，学習用文書において有害な文書に偏って出現する単語を有害キーワードとして情報量基準に基づき統計的に抽出し，キーワードが判定対象文書に含まれていれば有害として検出する．形態素解析を用いることなく判定が可能であり，判定ロジックも単純なキーワードマッチングであるため処理は高速であるが，精度に課題がある．複数のキーワードを組み合わせた判定や係り受け解析などを用いた深い言語解析を行うことで高精度化が可能となるが，高度な言語処理は処理時間が大きくなる．文献 3) の手法では，学習用文書と判定対象文書の特徴ベクトルをそれぞれ求め，判定対象文書の特徴ベクトルが学習用の有害文書の特徴ベクトルとどの程度類似しているかによって，判定対象文書の有害度合いを算出する．この手法では，判定対象文書に対して形態素解析を行う必要があるため，処理時間が大きくなることが課題である．

Web サイトの画像数やリンク数といった HTML に関連する特徴を用いて Web サイトの分類を行う手法も提案されている[4),5)]．文献 4) では，人手により Web サイトを観測することで，有害サイトの判定に役立つと思われる特徴を発見し，判定に利用する手法が提案されている．文献 5) も同様に，有害サイトの検出に役立つ特徴として画像数やリンク数などをあげ，リンク数が 10 以上のサイトは無害サイトに比べ有害サイトの方が多い，といった傾向を発見し，それらの特徴を組み合わせてベイジアンネットワークを用いて判定に利用している．しかし，これらの手法で抽出可能な特徴は観測者の主観や閲覧した Web サイトに依存するため，十分な性能を得ることが難しい．たとえば，著者らの予備実験において，有害サイトおよび無害サイト各 1 万サイトに対し，リンク数が 10 以上のサイトをすべて有害と判定したとすると，有害サイト全体の 75.7%を検出することができた（再現率 = 75.7%を意味する）が，有害と判定したサイトのうち，実際に有害であったサイトは 56.8%であり（適合率 = 56.8%を意味する），特徴量としての有効性は低いと考えられる．

Web サイトのハイパーリンクやソーシャルネットワークサービスの知り合い関係などを用いて Web サイトの分類を行う研究も報告されている[6),7)]．文献 6) では，ハイパーリンクの共起性とベクトル空間モデルを用いたクラスタを重ね合わせることにより，類似したクラスタを検出し，分類を実現している．文献 7) では，社会ネットワーク分析で用いられる指標を利用し，リンクに基づいてノードを高精度に分類する手法が提案されている．

3. 有害サイトの定義

本稿では，多様な有害サイトに対して高精度な判定を実現するため，実際の Web サイトを収集し，性能評価実験を実施する．サイトの収集においては，URL の収集と目視によるカテゴリ分類を行っている監視事業者からデータ提供を受けている．カテゴリ分類の一例を**表 1** に示す．有害・無害の判定は普遍的に一意に与えられるものではなく，ユーザが定義するポリシによって決定される．たとえば，アダルトや不法などのカテゴリに分類される Web サイトは多くのユーザにとって有害であるが，スポーツやゲームなどのカテゴリに分類される Web サイトもユーザが学校や企業にいる場合，本来従事すべき作業と関連が低いため，有害と定義する場合もある．本稿では，多くのユーザにとって有害であると考えられ

表 1　Web サイトのカテゴリ分類の詳細と有害・無害
Table 1　Detail categorization of Web pages and the definition of malicious/harmless categories.

カテゴリ名	内容	本稿における分類
セキュリティ・プロキシ	ハッキングや公開プロキシ	無害
出会い	出会い、結婚紹介	
金融	金融レート、投資商品、保険商品	
ギャンブル	ギャンブル一般、宝くじ、スポーツくじ	
ゲーム	オンラインゲーム、ゲーム一般	
ショッピング	オークション、通販、不動産	
コミュニケーション	ウェブチャット、メッセンジャー、掲示板	
ダウンロード	プログラムダウンロード、ストレージ	
職探し	転職・就職、キャリアアップ	
話題	イベント、話題	
成人嗜好	喫煙、飲酒	
オカルト	オカルト	
ライフスタイル	同性愛	
スポーツ	プロスポーツ	
旅行	観光情報、旅行商品、宿泊施設	
趣味	音楽、占い、芸能人、グルメ、娯楽一般	
宗教	伝統的な宗教、宗教一般	
政治活動・政党	政治活動、政党	
広告	広告、バナー、懸賞	
ニュース	ニュース一般	
不法	違法と思われる行為、薬物	有害
主張	テロリズム、武器、中傷、自殺	
アダルト	性行為、ヌード、性風俗、アダルトリンク	
グロテスク	グロテスク	
未承諾広告	迷惑メールに記載のリンク先	

る，不法，主張，アダルトなどを有害とし，その他のサイトについては無害と定義する．

6 章における性能評価実験などでは，収集した Web サイトのうち，学習用に 2 万サイト（有害・無害各 1 万サイト），判定用に 2 万サイト（有害・無害各 1 万サイト）を利用した．有害・無害サイトの各カテゴリにおける分布は監視事業者が管理しているデータ分布に従っており，実サービスに即した設定で性能評価を実施した．したがって，本稿における有害サイト検出の性能は，一般ユーザがインターネットを利用するシーンにおいても同程度であることが期待できる．

4. 提 案 手 法

4.1 提案手法の概要

提案手法における有害サイト検出処理の概要を**図 1** に示す．提案手法では，有害または無害のラベルが人手により付与された学習用サイトを利用した学習フェーズと，判定対象

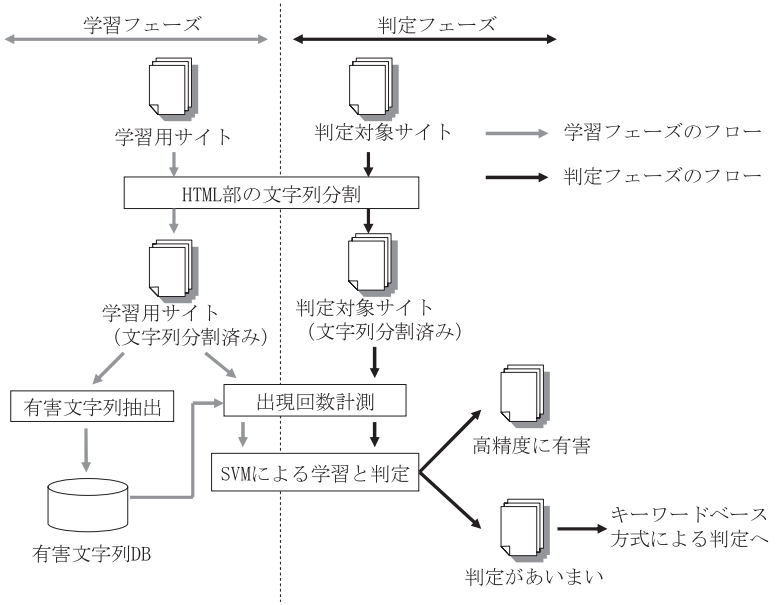

図 1　提案手法における処理フロー
Fig. 1　Overview of the processing flow in the proposed algorithms.

となるサイト集合から有害なサイトを検出する判定フェーズがある．学習フェーズでは初めに，有害サイトの HTML に偏って出現するような文字列を統計的な基準を用いて自動的に抽出する（4.2 節および 4.3 節）．次に，抽出した有害性の高い各文字列の学習用サイトにおける出現回数を特徴量として，SVM の学習を行う．判定フェーズでは学習フェーズと同様に，有害性の高い各文字列の判定対象サイトにおける出現回数を特徴量として SVM を用いて判定を行う（4.4 節）．

4.2　HTML 部の抽出と文字列分割

Web サイトから HTML 要素を抽出，文字列に分割する方法について説明する．ここで HTML 要素とは HTML ファイルから本文テキストを除いた <> で囲まれた部分とする．HTML ソースから本文テキストを抽出する方法については文献 8) や文献 9) などで提案されており，本稿では事前学習を必要とせず，計算量が少ないことを特徴とする文献 8) の手法を用いて，本文テキストと判定される部分を取り除いた HTML 要素を学習および判定に利用する．

次に，抽出した HTML 要素を文字列単位に分割する．区切り文字として，¥t , . / ! " = % & { } [ ] _ などを設定し，HTML 要素を分割する．**表 2** に HTML ソースと抽出した本文テキスト，本文テキストを除いた HTML 要素と HTML 要素を分割して抽出した文字列の例を示す．たとえば，<a href> タグからは a，href，http，www，ipsj（サーバ名），or，jp，10jigyo（フォルダ名やファイル名），html などが文字列として抽出される．

4.3　有害な文字列の抽出

学習用 Web サイトにおいて有害サイトの HTML 部に偏って出現する文字列を自動的に抽出する．抽出手法として文献 2) と同様の手法を用いる．文献 2) では，ある文字列 s が有害なサイトに偏って出現する度合いを表す指標 $E(s)$ を AIC（赤池情報量基準）[10] を用いて算出する．**表 3** のように，ある文字列 s が出現する有害サイト数 N_{11} と無害サイト数 N_{21}，文字列 s が出現しない有害サイト数 N_{12} と無害サイト数 N_{22} の 4 つの値を学習用サイトに出現するすべての文字列について求める．文献 2) では文字列 s が有害な文書に偏って出現する度合い $E(s)$ を文献 11) の知見をもとに，AIC の独立モデルに対する値 AIC_IM および従属モデルに対する値 AIC_DM を用いて，次のように定義している．

$$
\begin{aligned}
&N_{11}(s)/N(s) > N_{12}(s)/N(\neg s) \text{ のとき,}\\
&\quad E(s) = AIC\_IM(s) - AIC\_DM(s)\\
&N_{11}(s)/N(s) \leq N_{12}(s)/N(\neg s) \text{ のとき,}\\
&\quad E(s) = AIC\_DM(s) - AIC\_IM(s) \qquad (1)
\end{aligned}
$$

表 2　HTML 要素の抽出と文字列分割の例
Table 2　Example of extraction and parsing of HTML elements.

HTMLソース

<td></td> </tr> <tr> <td height="80" valign="top" class="font_glay_11">電子情報通信学会情報・システムソサイエティ（ISS）及びヒューマンコミュニケーショングループ（HCG)と情報処理学会（IPSJ)の合同で開催致します本フォーラムは，IPSJ全国大会とISSソサイエティ大会との流れを汲むものですが，従来の大会の形式にとらわれずに，新しい発表形式を導入し，タイムリーな情報発信，活気ある議論・討論，多彩な企画，他分野研究者との交流などを実現してゆきたいと考えております．
 ※FIT創設の経緯とIPSJ-ISS覚書</td>

本文テキスト

電子情報通信学会情報・システムソサイエティ（ISS）及びヒューマンコミュニケーショングループ（HCG)と情報処理学会（IPSJ)の合同で開催致します本フォーラムは，IPSJ全国大会とISSソサイエティ大会との流れを汲むものですが，従来の大会の形式にとらわれずに，新しい発表形式を導入し，タイムリーな情報発信，活気ある議論・討論，多彩な企画，他分野研究者との交流などを実現してゆきたいと考えております．※FIT創設の経緯とIPSJ-ISS覚書

本文テキストを除いたHTML要素

<td></td> </tr> <tr> <td height="80" valign="top" class="font_glay_11">
 ※</td>

HTML要素を分割した文字列(括弧内は複数出現回数)

a(2), alt, blank, br, class, fit(2), font, found, gaiyo, gif, glay, height(2), href, html, http, img(2), ipsj, 10jigyo, jp, or, src, target, td(4), top, tr(2), valign, width, www

表 3　$E(s)$ 値算出に用いる文字列 s の出現回数
Table 3　Appearance frequency of string s to evaluate $E(s)$ value.

	文字列 s が出現	文字列 s が非出現	合計
有害サイト	$N_{11}(s)$	$N_{12}(s)$	N_p
無害サイト	$N_{21}(s)$	$N_{22}(s)$	N_n
合計	N(s)	N(¬s)	N

表 4　文字列の出現回数と $E(s)$ 値の例
Table 4　Example of $E(s)$ values and appearance frequency.

文字列	$N_{11}(s)$	$N_{12}(s)$	$N_{21}(s)$	$N_{22}(s)$	E(s)
S_1	100	1000	50	9850	122.9
S_2	10	1090	900	9000	-55.6
S_3	100	1000	900	9000	-2.0

ここで，$AIC\_IM(s)$，$AIC\_DM(s)$ はそれぞれ文献 10) の定義に従って，次の式で与えられる．

$$
\begin{aligned}
&AIC\_IM(s) = -2 \times MLL\_IM + 2 \times 2 \\
&MLL\_IM = N_p(s)\log N_p(s) + N(s)\log N(s) + N_n(s)\log N_n(s) \\
&\qquad + N(\neg s)\log N(\neg s) - 2N\log N \\
&AIC\_DM(s) = -2 \times MLL\_DM + 2 \times 3 \\
&MLL\_DM = N_{11}(s)\log N_{11}(s) + N_{12}(s)\log N_{12}(s) + N_{21}(s)\log N_{21}(s) \\
&\qquad + N_{22}(s)\log N_{22}(s) - N\log N
\end{aligned}
\tag{2}
$$

具体例として，有害サイトに偏って出現する文字列 S_1 と無害サイトに偏って出現する文字列 S_2，偏りなく出現する文字列 S_3 の具体的な出現回数と $E(s)$ 値の例を**表 4** に示す．

S_1 は有害サイトに偏って出現する文字列であるため，有害度合いを表す指標 $E(s)$ が正の値をとり，S_2 は無害サイトに偏っているため $E(s)$ は負の値をとる．S_3 は偏りなく出現するため，$E(s)$ は 0 に近い値となる（この例では，AIC の独立モデルを用いるか，従属モデルを用いるかの違いにより -2.0 の差が生じる）．この手法により，有害性の高いリンク先のサーバ名や有害サイトで頻繁に用いられるポップアップなど Web ブラウザに特定の動作を要求する javascript 関数名などを自動的に抽出することができる．

実際の Web サイトから抽出された具体的な文字列の一例として “writeflash” を紹介する．多くのブラウザでは，Web サイトのリンクにマウスカーソルを合わせると，リンク先の URL がブラウザ下部に表示されるが，“writeflash” はこのリンク先の URL を表示させない目的で利用される javascript 関数の一部である．リンク先を伏せることで，ユーザが目的のリンク先を発見する過程において，より多くのアフィリエイトリンクを参照させる効果があるが，有害サイト以外ではこのような仕組みは意味をなさないためほとんど見られず，有害サイトに特徴的な文字列といえる．

抽出した有害サイトの検出に役立つ各文字列について単独での性能を評価するための予備実験として，各文字列を K 回以上含む Web サイトを有害と判定する方式において，K の値を変化させたときの再現率（Recall）と適合率（Precision）の関係を**図 2** に示す．実験データとしては，3 章に記載した学習用の Web サイト（有害・無害各 1 万サイト）を利用し，提案手法により統計的に抽出された有害性の高い文字列（ここでは X，Y とする）と，文献 4) や文献 5) で有効とされている人手により観測された特徴量である画像数，リンク数をそれぞれ単独で用いて有害サイトを検出したときの性能を比較する．

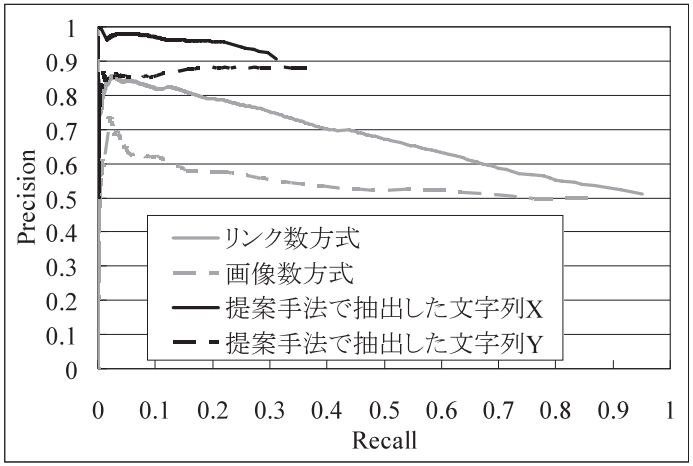

図 2　提案手法により HTML 部から抽出された文字列 X，Y と画像数，リンク数をそれぞれ K 回以上含むサイトを有害と判定した際の性能比較
Fig. 2　Comparison of the performance of X and Y, which are extracted by the proposed algorithms, the number of links, and the number of image files whereby Web pages that contain each string more than K times are judged as malicious.

本稿では有害サイト検出の再現率と適合率を判定対象となる Web サイト集合中の全有害サイト数 All（本実験では 1 万サイト），各手法で有害と判定したサイト数 Judge，有害と判定したうち，正しく有害と判定できたサイト数 Correct を用いて，次のように定義する．

$$Recall = Correct/All \tag{3}$$

$$Precision = Correct/Judge \tag{4}$$

図 2 から提案手法により得られた文字列 X，Y に基づく方式では同じ再現率においては，画像数やリンク数に基づく方式と比べて適合率が大きい傾向にあることが分かる．X，Y に基づく方式の再現率の最大値（1 回以上文字列が出現する有害な Web サイトの割合）は画

像数やリンク数に基づく方式と比べて低いが，これは複数の文字列を組み合わせることで向上させることができる．このように適合率の高い特徴を持つ文字列を組み合わせることにより，提案手法では高精度を実現することが可能となる．一方，文献 4) や 5) であげられる人手による観測で有効性が高いとした画像数やリンク数などに基づく方式は適合率が低いため，組み合わせると全体の適合率が低下することや，利用する識別器のパラメータの最適化が複雑になり，未知データの識別に対する汎化性能が低下するといった問題が生じる．

4.4　SVM による学習と判定

4.3 節で抽出した有害サイトの検出に役立つ文字列を組み合わせて SVM[12] を用いて有害サイトの特徴を学習し，検出する．具体的には，抽出した文字列 $S_1, S_2, S_3, \cdots, S_m$ と各サイトにおける各文字列の出現回数 $N_1, N_2, N_3, \cdots, N_m$ からなる行列を SVM の入力として与える．学習フェーズでは加えて各サイトが有害または無害を表すラベル Label も合わせて与えることで SVM を学習させる．**表 5** に SVM の入力例を示す．

有害サイトの検出に SVM を用いることの妥当性について述べる．本手法の利用シーンを考慮すると，学習データに対して正しい識別ができることよりも判定対象データ（未知のデータ）に対して汎化性能を示す識別器を利用することが望ましい．SVM は一般に汎化性能に優れているといわれており，本手法に適切と考えられる．予備実験として，提案手法の識別器として SVM と決定木を用いた識別器である C4.5[13] を用いた場合の性能を比較評価した．学習データとして 3 章で紹介した学習用 Web サイト 2 万サイト（有害，無害各 1 万サイト）を用いて SVM と C4.5 をそれぞれ学習させ，判定用の Web サイトとは異なる 2 万サイト（有害，無害各 1 万サイト）を判定し，F 値について評価した．SVM を用いた場合の F 値は 69.1%，C4.5 を用いた場合の F 値は 59.4%であり，汎化性能の高い SVM の方が本手法に適していることが期待される．C4.5 は著名な識別器であるが，このほかに Neural Network[14] や Bayesian Filtering[15] なども有効性があると考えられ，これらを利用した際の性能の検証は今後の課題である．

表 5　SVM の入力となる特徴量の例
Table 5　Example of SVM features.

	S_1	S_2	S_3	…	S_m	Label (学習データのみ)
サイト 1	N_{11}	N_{12}	N_{13}	…	N_{1m}	1
サイト 2	N_{21}	N_{22}	N_{23}	…	N_{2m}	0
…	…	…	…	…	…	…
サイト X	N_{X1}	N_{X2}	N_{X3}	…	N_{Xm}	0

また，SVM では判定の信頼度を計算することが可能であり，有害または無害と判定する閾値をそれぞれ設定することが可能である．閾値を高く設定すれば再現率は低いが適合率は高くなる．閾値を低く設定すれば再現率は高くなるが，適合率は低くなる．6 章における実験では閾値を変化させたときの提案手法の再現率，適合率の関係を示す．

5.　提案手法とキーワードベース方式の併用

提案手法は Web サイトの HTML 部分のみを用いて判定を行うため，本文を対象として判定を行う既存のキーワードベース方式と組み合わせて利用することで，さらに高精度な判定を行うことが可能と考えられる．提案手法と従来手法[2] によって検出可能な有害サイトの相関関係を調べるための予備実験を行った．

4.4 節における実験と同様に学習用 Web サイト 2 万サイト，判定用 Web サイト 2 万サイトを用いた．提案手法，従来手法それぞれにおいて，再現率が $10, 20, 30, \cdots, 90$ (%) のとき，(1) 提案手法でのみ有害と判定したサイト数，(2) 従来手法でのみ有害と判定したサイト数，(3) 両方の手法で有害と判定したサイト数を**図 3**，**図 4** に示す．再現率が大きくなる

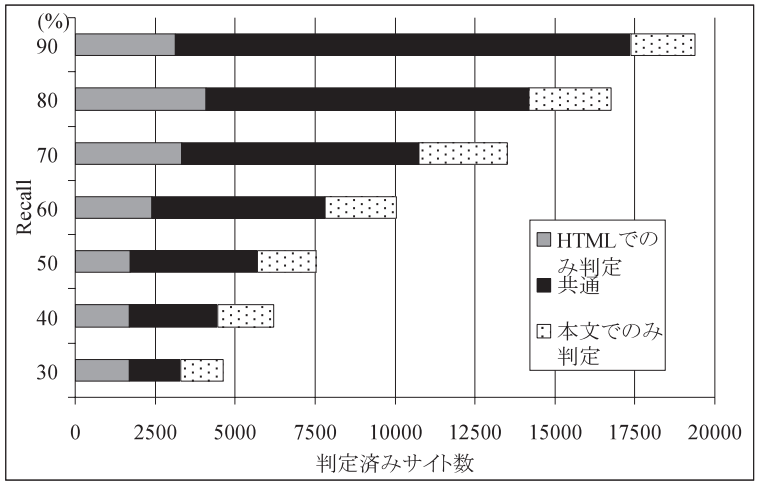

図 3　提案手法と従来手法において有害と判定したサイト数
Fig. 3　Number of Web pages detected as malicious by the proposed algorithms and by the existing algorithms.

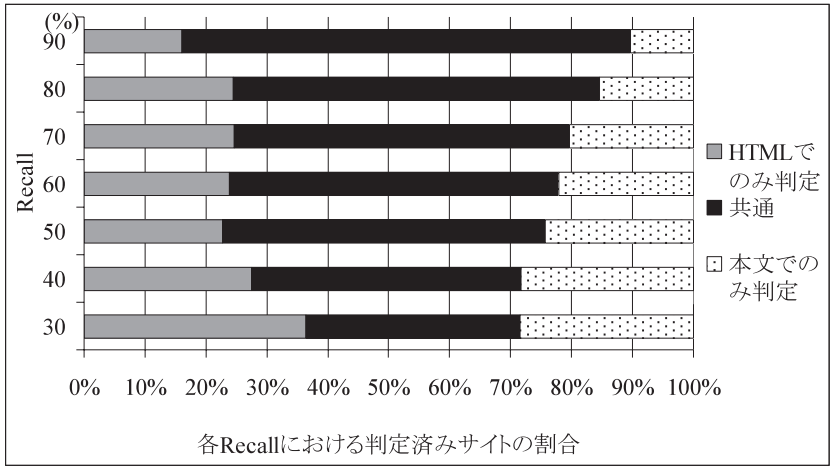

図 4　提案手法と従来手法において有害と判定したサイト数（割合）
Fig. 4　Ratio of Web pages detected as malicious by the proposed algorithms and by the existing algorithms.

に従って，両方の手法で共通に有害と判定したサイトの割合が増加するが，再現率 90%においても，各手法でのみ判定可能なサイトが存在することが分かる．この結果から提案手法と従来手法を組み合わせて利用することで，より多くの有害サイトを検出することが可能と考えられる．6 章における実験では，提案手法において有害と判定する SVM の信頼度閾値を高めに設定し，明らかに有害なサイトを検出し，閾値に満たない判定があいまいであるようなサイトについてはキーワードベース方式を用いて判定するという手法の性能についても評価を実施する．

6. 性能評価実験

6.1 提案手法と従来手法の性能比較評価

提案手法を実装し，キーワードベース方式の従来手法[2] との性能比較評価実験を実施した．実験環境と実験手順を下記に示す．

実験環境：計算機（1core 2.53 GHz，64 GB RAM，Linux OS），提案手法で利用する SVM として Lib SVM [16]，従来手法で学習時に利用する形態素解析器として MeCab [17] を用いた．また提案手法，従来手法の実装には C 言語を用いた．

利用データ：提案手法と従来手法の比較には，3 章で定義した Web サイト 4 万サイト（学習用サイト，判定対象サイト各 2 万サイト，有害と無害の比率は 1 : 1）を用いた．

評価指標：提案手法，従来手法において再現率と適合率を評価する．また，各手法において 1 サイトの判定に要する平均処理時間についてもあわせて評価する．

実験手順：次にあげる 5 つの手法の性能を比較評価する．（手法 1）提案手法単独，（手法 2）従来手法単独，（手法 3）提案手法において判定の信頼度が閾値以上のサイトについては有害と判定し，閾値以下の判定があいまいであるサイトについては従来手法を用いて判定する手法（以降では複合手法と呼ぶ），（手法 4）従来手法では単純な文字列一致により有害サイトを検出するが，提案手法と同様に SVM を用いて有害性の高い単語を組み合わせて判定する手法（以降では従来 + SVM 手法と呼ぶ）．手法 1 では，HTML から抽出した文字列 26 個を利用した．手法 2 ではテキスト本文から抽出した単語 25,000 個を利用した．手法 4 については提案手法と同量の 26 個の単語を利用した場合（手法 4-a）と，10,000 個の単語を利用した場合（手法 4-b）についてそれぞれ評価した．

実験結果：各手法における再現率と適合率の関係を**図 5** に示す．(1) の提案手法と (2) の従来手法を比較すると，提案手法は 26 個という少数の文字列のみを利用したにもかかわらず，再現率 50%以下の領域においては適合率が 90%以上ときわめて高い適合率を実現している．これはテキスト部で利用される単語の種類数に比べ，HTML 要素の種類数が少ない点や，有害サイトは類似した構成を持ちやすい傾向にあり，HTML 要素にそれらの特徴が多く現れているためと考えられる．再現率の高い領域においては従来手法の方が適合率は高くなる傾向が確認できるが，提案手法において有効性の高い文字列をさらに追加することで適合率，再現率の向上が期待される．

(3) の複合手法では再現率が 50%となるまで (1) の提案手法を用いて判定を行い，未判定のサイトを (2) の従来手法を用いて判定した．(1) の手法において性能が低下する再現率が高い領域においても性能が改善され，従来手法と比べてすべての再現率において高い適合率を実現することが分かった．特に再現率 70%においては従来手法と比べて適合率が 68.8%から 78.1%に 9.3 ポイント向上するなど，きわめて効果的であることが分かった．このとき，F 値では 69.4%から 73.8%に向上する．提案手法は HTML 部分を対象とし，従来手法と異なる情報を利用していることから組み合わせて判定を行うことで高精度化の可能性があるという 5 章の仮説を確認できたといえる．

(4-a) の手法は (1) と同数の 26 個の単語をテキスト部から抽出したが，(1) よりも全体的に低い性能となった．これはテキスト部で利用される単語の種類数に比べて HTML 要素の

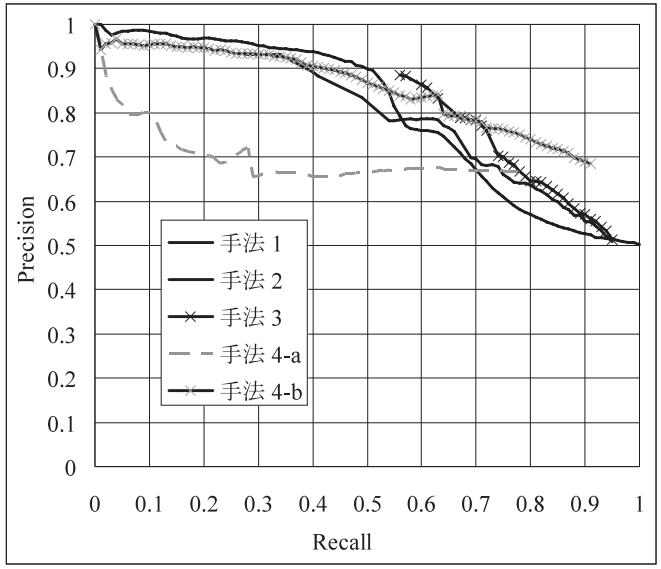

図 5　各手法における再現率，適合率の比較
Fig. 5　Comparison of the performance of each algorithm.

種類数が少ないため，提案方式ではキーワードベースの方式に比べてより少ない特徴量で高精度な判定が実現できるためと考えられる．(4-b) の 10,000 単語を組み合わせて SVM を用いて判定する手法は再現率の高い領域において (1) の提案手法や (3) の複合手法よりも性能が高いことが分かった．

次に，判定に要した処理時間を**表 6** に示す．(1) の提案手法と (2) の従来手法の処理時間はそれぞれ 3.85 msec，3.57 msec とほぼ同程度の処理時間となった．これは形態素解析のみを行った場合の処理時間と比べて半分程度であり，文献 3) のような高度な言語解析を行うキーワードベース方式と比べて高速であるといえる．また，(4-b) の手法は再現率の高い領域において (1) の提案手法や (3) の複合手法よりも性能が高いが，多数の特徴量を組み合わせて判定を行うため，処理時間が大きくなる点が課題である．提案手法は少数の文字列でも比較的高精度を実現できるため，これらの問題を解決することができる点でも実用的である．

表 6　判定に要した処理時間の比較
Table 6　Comparison of the processing time of each algorithm.

	1 サイトの判定に要した平均処理時間(msec)
手法 1(提案)	3.85
手法 2(従来)	3.57
手法 3(提案+従来の複合)	3.65
手法 4-a(従来+SVM 26 単語)	3.50
手法 4-b(従来+SVM 10,000 単語)	12.12
形態素解析のみ(参考)	6.82

6.2　日本語サイトと他言語サイトの判定性能比較

提案手法は Web サイトの本文情報を判定に利用しないため，他言語のサイトに対しても適用可能であることが期待されるため，他言語サイトと日本語サイトの判定性能比較実験を実施した．他言語サイトの性能評価には，2,000 サイト（有害・無害各 1,000 サイト）を利用した．他言語サイトであるかどうかを判定する方法として，日本語を表す文字コードを含んでおらず，かつ 50 Byte 以上の本文テキストを含んでいるサイトを利用した．利用したサイトの主な言語としては英語，ヨーロッパ圏の言語，中国語，韓国語，ロシア語などである．

日本語サイトから学習した特徴をもとに，他言語サイトを判定した際の性能と日本語サイトを判定した際の性能とを比較したグラフを**図 6** に示す．日本語サイトを判定した際の性能と比較すると，Recall の小さい領域において Precision の若干の低下が見られるものの，Recall が 50%以降の Precision は日本語サイトと同程度であることが分かり，提案手法は他言語のサイトに対しても一定の有効性が見られる．Precision が低下した理由として，日本語の有害サイトの多くが利用するアフィリエイトサーバが他言語のサイトでは必ずしも利用されていないことが原因と考えられる．

7. ま と め

本稿では高速かつ高精度に有害サイトを検出するため，Web サイトの HTML を対象とした有害サイト検出手法を提案した．提案手法では有害サイトの HTML に偏って出現するような文字列を情報量基準に基づき統計的に抽出し，SVM を用いてこれらの特徴を組み合わせて有害サイトの検出を行う．提案手法は Web サイトの本文の情報を利用しないため，既存のキーワードベース方式によって検出が困難なサイトも検出が可能であることを，各手

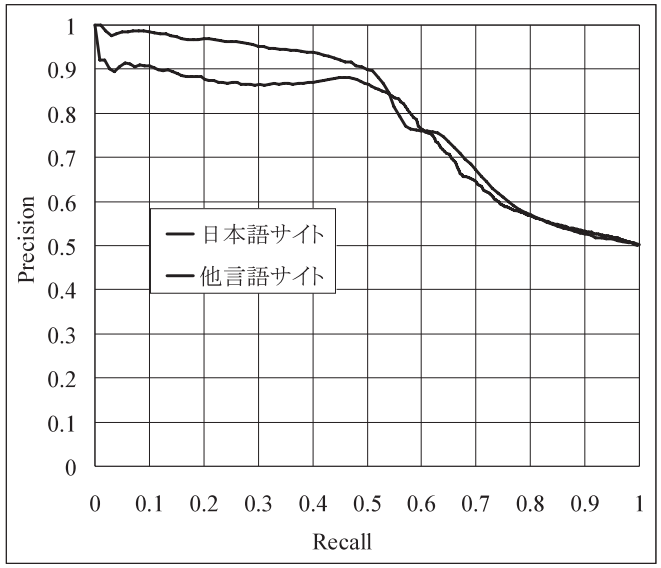

図 6　他言語サイトと日本語サイトに対する提案手法の性能比較
Fig. 6　Performance comparison of the classification of foreign and Japanese Web pages.

法で有害と判定するサイトの相関から検証した．性能評価実験においては，提案手法単体で利用した場合，再現率 50.0%，適合率 90.3%ときわめて高い適合率が実現できることを確認し，さらに既存のキーワードベース方式と提案手法を組み合わせて判定を行う複合手法では，再現率 70.0%，適合率 78.1%を達成した．これは従来のキーワードベース方式の同程度の再現率における適合率と比較して 9.3 ポイント向上しており，きわめて高性能なフィルタリングシステムを実現したといえる．

今後の課題として，本稿では識別器として SVM を用いたが，Neural Network や Bayesian Filtering など提案手法の親和性の評価や，海外のサイトにおいてサーバ名の違いなどによる適合率の低下が見られた点について，改善手法の検討などを予定している．

謝辞　本研究は，（独）情報通信研究機構の委託研究「高度通信・放送研究開発委託研究/インターネット上の違法・有害情報の検出技術の研究開発」の一環として実施した．また，日頃御指導いただく KDDI 研究所中島康之所長，鈴木正敏所長に深謝いたします．

参 考 文 献

1) 総務省：ブログの実態に関する調査研究 (2008). http://www.soumu.go.jp/iicp/chousakenkyu/seika/ houkoku.html#2008
2) 柳原　正，松本一則，小野智弘，滝嶋康弘：トピック判定における n-gram の組み合わせ手法の検討，第 7 回情報科学技術フォーラム（FIT2008）論文集 (2008).
3) 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.84, No.6, pp.1158–1166 (2001).
4) 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).
5) Ho, W.H. and Watters, P.A.: Statistical and Structural Approaches to Filtering Internet Pornography, *Proc. IEEE International Conference on Systems, Man and Cybernetics*, pp.4792–4798 (2004).
6) 高橋　功，三浦孝夫：ハイパーリンクの共起性を用いたクラスタリング手法，*DEWS2005*, 1C-i12 (2005).
7) 唐門　準，松尾　豊，石塚　満：リンクに基づく分類のためのネットワーク構造を用いた属性生成，情報処理学会論文誌，Vol.49, No.6, pp.2212–2223 (2008).
8) 吉田光男，山本幹雄：教師情報を必要としないニュースページ群からのコンテンツ自動抽出，日本データベース学会論文誌，Vol.8, No.1, pp.29–34 (2009).
9) Lin, S.H. and Ho, J.M.: Discovering Informative Content Blocks from Web Documents, *Proc. ACM SIGKDD*, pp.588–593 (2002).
10) 鈴木義一郎：情報量基準による統計解析入門，（株）講談社サイエンティフィク（編），pp.80–96，（株）講談社，東京 (1995).
11) Matsumoto, K. and Hashimoto, K.: Schema Design for Causal Law Mining from Incomplete Database, *Proc. Discovery Science: 2nd International Conference* (*DS'99*), pp.92–102 (1999).
12) Cortes, C. and Vapnik, V.: Support-Vector Networks, *Machine Learning*, pp.273–297 (1995).
13) Quinlan, J.R.: *C4.5: programs for machine learning*, Morgan Kaufmann (1993).
14) Haykin, S.: *Neural Networks: A Comprehensive Foundation*, Prentice Hall PTR (1998).
15) Hand, D.J., Mannila, H. and Smyth, P.: *Principles of Data Mining*, The MIT Press (2001).
16) Fan, R., Chen, P. and Lin, C.: Working set selection using the second order information for training SVM, *Journal of Machine Learning Research*, Vol.6, pp.1889–1918 (2005). http://www.csie.ntu.edu.tw/˜cjlin/libsvm/
17) Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying conditional random fields to japanese morphological analysis, *Proc. 2004 Conference on Empirical Methods*

in Natural Language Processing (*EMNLP-2004*), pp.230–237 (2004).
http://mecab.sourceforge.net/

(平成 22 年 9 月 30 日受付)
(平成 23 年 5 月 14 日採録)

推　薦　文

違法・有害サイト検出に関して，html のタグに偏って出現する文字列を利用した，高速かつ高精度で有用性の高い検出手法の提案を行っている．評価実験により，既存のキーワードベース方式と比較し，適合率 9.3%向上などの性能向上を確認している．以下の理由により FIT2010 ベストペーパー賞にふさわしいと判断し推薦する．

1. HTML のタグに着目している点に一定の独創性が認められる．
2. 大規模な評価実験により，提案手法の優位性が実際に検証されている．
3. 他の手法との組合せが容易であり，さらなる精度向上や他分野への応用が可能である．

（FIT2010 第 9 回情報科学技術フォーラム　プログラム委員長　村山優子）

池田　和史（正会員）

2006 年大阪大学基礎工学部情報科学科飛び級のため中退．2008 年同大学大学院博士前期課程修了．同年 KDDI（株）入社，研究所所属．自然言語処理等の研究に従事．電子情報通信学会，日本データベース学会各会員．

柳原　　正（正会員）

2002 年慶應義塾大学環境情報学部卒業．2004 年同大学大学院修士課程修了．2005 年 KDDI（株）入社，研究所所属．リコメンダシステム，テキストマイニング等の研究に従事．電子情報通信学会，日本データベース学会各会員．

服部　　元（正会員）

1973 年生．1996 年神戸大学工学部電気電子工学科卒業．1998 年同大学大学院自然科学研究科電気電子工学専攻修士課程修了．同年国際電信電話（株）(現，KDDI（株)）入社．現在，（株）KDDI 研究所テキスト情報処理グループ研究員．この間，ネットワーク管理，ITS，ソフトウェアエージェントの研究に従事．電子情報通信学会会員．

松本　一則

1984 年京都大学工学部情報工学科卒業．1986 年同大学大学院修士課程修了．同年国際電信電話（株）入社，研究所所属．現在，KDDI 研究所知能メディアグループにて，マルチメディア検索，コンテンツ配信の研究開発に従事．電子情報通信学会，日本データベース学会各会員．工学博士．

小野　智弘

1992 年慶應義塾大学理工学部電気工学科卒業．1994 年同大学大学院理工学研究科計算機科学専攻修士課程修了．同年国際電信電話（株)(現，KDDI（株)）入社．1999〜2000 年スタンフォード大学電気工学科客員研究員．現在，（株）KDDI 研究所にて，データマイニング，リコメンダシステムの研究に従事．1996 年度情報処理学会学術奨励賞受賞．電子情報通信学会会員．工学博士．

滝嶋　康弘

1986 年東京大学工学部電気工学科卒業．1988 年同大学大学院電子工学専攻修士課程修了．同年国際電信電話（株)(現，KDDI（株)）入社．現在，（株）KDDI 研究所知能メディアグループリーダ．この間，動画像の符号化方式，動画通信システム，情報理論の研究・開発に従事．電子情報通信学会，映像情報メディア学会，画像電子学会各会員．工学博士．