Vol.2010-DPS-142 No.44
Vol.2010-CSEC-48 No.44
2010/3/5

コンテンツベースフィッシング検知手法の大規模実例評価と改良

加藤 慧[†]　小宮山 功一朗[††]　瀬古 敏智[††]
一瀬 友祐[††]　河野 耕平[†]　吉浦 裕[†]

概要 金融機関等の Web サイトに成りすまして個人情報を詐取するフィッシング詐欺が社会問題になっている．コンテンツベース（CB）のフィッシング検知方式は，検査対象サイトのテキストからキーワードを抽出し，これを用いた Web 検索によって正規サイトを求め，検査対象サイトと比較することでフィッシング判定を行う方式である．ブラックリスト方式等で必要となるリストの管理が不要であり，また，検知率が高いとされている．しかし，大量の実例データを用い CB 方式を評価した例はなく，その性能について明確な知見は存在しない．また，従来の CB 方式は，英語サイトにのみ対応していた．そこで，日英両言語の CB システムを実装し，JPCERT/CC の保有する 843 件のフィッシングサイトを用いて評価した． 843 件全てについて正しいフィッシング判定が得られた．そのうち 697 件については正規サイトの検索に成功し，CB 方式の想定通りの動作であった．残りの 146 件については，正規サイトが検索されない理由を分析した．

Content-based phishing detection methods improvements tested against actual phishing incidents

Kei Kato[†]　Koichiro Komiyama[††]　Toshinori Seko[††]
Yusuke Ichinose[††]　Kohei Kawano[†]　Hiroshi Yoshiura[†]

Abstract Phishing attacks steal personal information by impersonating legitimate sites and convincing users to send personal information. The Content-Based (CB) phishing detection method extracts keywords from text of a suspect site and uses these keywords to determine the legitimate sites that the suspect site may impersonate. The CB method does not require the lists of sites that are necessary in conventional methods such as black-lists. However, the CB method has never been evaluated against a significant number of real phishing sites; therefore, its performance in real situations has not been proven. It has previously been used only for English sites, so its performance for non-English sites has been untested. We have developed a CB system that can be used for both English and Japanese sites and have evaluated it using 843 real phishing sites that have been captured by JPCERT/CC.

1. はじめに

近年，インターネットの急速な普及によって，子供や高齢者などコンピュータリテラシーの低いユーザによるインターネットの利用が一般化してきた．これに伴って，コンピュータリテラシーの低いユーザをターゲットとしたフィッシング詐欺が急増している．フィッシング詐欺とは，金融機関や公的機関，ソーシャル・ネットワーキング・サービス（SNS）等を装った偽のウェブサイト（フィッシングサイト）を制作し，そこからユーザの個人情報を詐取する詐欺の総称である．フィッシング詐欺による被害額は，2006 年度の全米被害額が 28 億ドル，2007 年度では 32 億ドルと年々増加している[1]．また，従来のフィッシング詐欺の多くは，米国を中心とした国外でのものであったが，最近では日本国内でのフィッシング詐欺も増加しており，2010 年 1 月には国内での逮捕者も出ている．

フィッシング詐欺への対策方法として，様々なフィッシング検知方式が提案されている．その中でも，Yueら[2]，中山ら[3]によって提案されているコンテンツベース方式は，データベースのメンテナンスが不要で，かつ即時性の高いフィッシング検知方式として注目されている．この方式は，フィッシングサイトが正規サイトの模倣であることに注目し，検索エンジンを利用して正規サイトを探し出すことで，フィッシング詐欺検知を行う方式である．

しかし，この方式には次の問題がある．

- 検知性能について，小規模な評価しか行われておらず，大規模な実例データを利用した評価が行われていない．
- 適用可能な言語として，英語に対応した実装評価は行われているが，日本語のフィッシングサイトに対応する実装評価が行われていない．
- 原理上の問題として，文字情報が少ないウェブページの検査が行えない．
- 実装上の問題として，HTML パース方法および特殊文字の扱い方が検知性能に与える影響について十分に検討されていない．

そこで本稿では，大量のフィッシング実例データを用いたコンテンツベース方式の評価ならびに日本語のフィッシングサイトに対応したシステムの実装評価を行う．実験には，国内のセキュリティインシデント対応機関である JPCERT コーディネーションセンターが保有する 843 件のフィッシングサイト実例データを用いて行う．

[†] 電気通信大学
The University of Electro-Communications
[††] JPCERT コーディネーションセンター
Japan Computer Emergency Response Team Coordination Center

Vol.2010-DPS-142 No.44
Vol.2010-CSEC-48 No.44
2010/3/5

2. 従来方式

従来のフィッシング検知手法として，ホワイトリストやブラックリスト等のデータベースを用いた方式がある．

- **ホワイトリスト方式** - 正規サイトを記録したホワイトリストと比較し，載っていないウェブサイトを信頼できないと判断する[4]．この方式では，中小企業や新規サイトをすべて網羅することは難しく，ホワイトリストに載っていないサイト以外はフィッシングサイト扱いされるという可能性がある．
- **ブラックリスト方式** - フィッシングサイトを記録したブラックリストと比較し，載っていたサイトを信頼できないと判断する[5]．ブラックリストは，フィッシングサイトを見た人がブラックリストの管理組織に通報して，はじめて登録される．そのため，フィッシングサイトが現れてから，実際にブラックリストに登録されるまでには時間差が存在する．したがって，ブラックリストに登録されるまでのタイムラグの間に，閲覧してしまったユーザを守ることはできない．

これらの方式は，いずれもデータベースの頻繁なメンテナンスが必要とされるため，管理コストの高さや，即時性の高い判断ができないことが問題となる．そこで，データベースの更新が不要で，かつ即時性の高いフィッシング検知が可能なコンテンツベース方式が提案されている．この方式では，フィッシングサイトが正規サイトの模倣であることに注目し，検索エンジンを利用して正規サイトを探し出すことで，フィッシング検知を行う方式である．その詳細は次章で述べる．

3. コンテンツベース方式

3.1 方式

コンテンツベース方式とは，コンテンツの類似性を利用したフィッシング検知方式である．フィッシングサイトは，ユーザを騙すために特定のウェブサイトになりすます．このようなフィッシングサイトの多くは，正規サイトのコンテンツをコピーまたは模倣して作成されたものである．そのため，フィッシングサイトと正規サイトの内容は酷似しており，そこに出現する言葉や見た目には同じ特徴が見られる．コンテンツベース方式では，このような類似性に着目することで，フィッシング検知を行う．

コンテンツベース方式による処理は次の通りである（図 1）．

- 検査対象ページ内から特徴的な語句を抽出する．
- 特徴度の高い上位 N 件の単語をキーワードとしてウェブ検索を行う．
- もし検査対象ページのドメインが検索結果の上位 M 件の中に含まれていれば，正規サイトと判断する．含まれていなければ，フィッシングサイトと判断する．

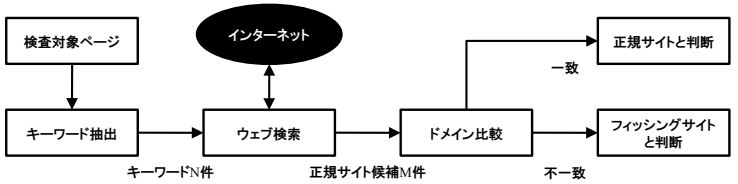

図 1 コンテンツベース方式

この方式のフィッシング検知方法が有効である根拠として，ウェブ検索エンジンの特性を説明する．特徴的な語句によってウェブ検索をした結果の中には，模倣元である正規サイトが現れる一方で，フィッシングサイトは現れない．なぜなら，フィッシングサイトは平均存続期間が 3.1 日と短く[6]，また他のウェブサイトからリンクされることも稀なため，検索エンジンからの評価が低いからである．そして，一般的な企業の正規サイトは，これとは逆の性質を有しており，検索エンジンからの評価が高い．すなわち，コンテンツベース方式は，検索エンジンの特性を利用することでホワイトリストを動的に生成しているとも表現できる．

3.2 評価基準

フィッシング検知方式の評価は，次の点について行われる．

- **フィッシング検知率** - フィッシングサイトを検査し，フィッシングと正しく判断した率
- **正規サイト誤検知率** - 正規サイトを検査し，フィッシングと誤って判断した率

既存研究では，様々な誤検知防止手法が提案されているが，ここではそれらの手法を用いない基本的コンテンツベース方式における実験方法および結果を説明する．

3.3 実験方法

Yue らは，次の方法でそれぞれ実験および評価を行っている．

- **Basic TF-IDF** - TF-IDF 上位 5 件の単語をキーワードにウェブ検索を行い，検索結果上位 30 件のドメインと比較する．検索結果が 0 件の場合，判断不可能とする．
- **Basic TF-IDF + domain** - TF-IDF 上位 5 件の単語に検査対象ページのドメイン名を加えたものをキーワードとしてウェブ検索を行い，検索結果上位 30 件のドメインと比較する．
- **Basic TF-IDF + ZMP** - Basic TF-IDF と同様の実験を行う．検索結果が 0 件の場合，フィッシングサイトであると判断する．（ZMP : Zero Means Phishing）
- **Basic TF-IDF + domain + ZMP** - Basic TF-IDF + domain と同様の実験を行う．検索結果が 0 件の場合，フィッシングサイトであると判断する．

Vol.2010-DPS-142 No.44
Vol.2010-CSEC-48 No.44
2010/3/5

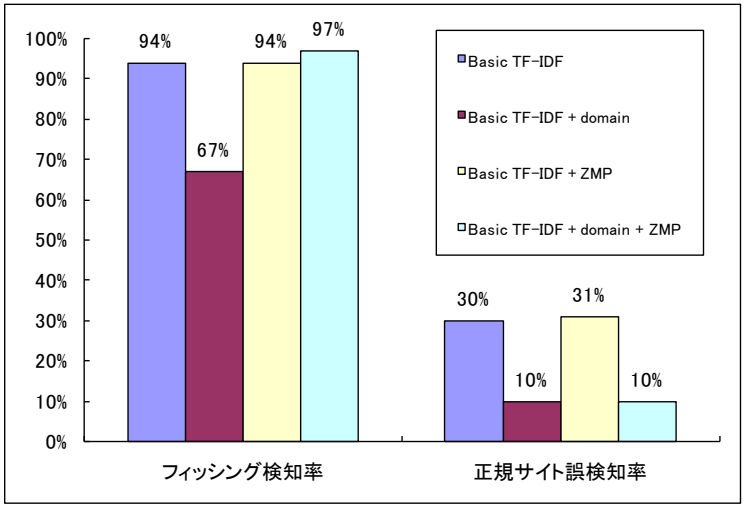

図 2 既存研究による評価

Yue らの実験で使用したサンプルは，フィッシングサイト 100 件，正規サイト 100 件であり，いずれも全て英語のウェブページである．これらによる実験結果を図 2 に示す．

4. 実装

Yueらの方式に従い，英語のウェブページに適用可能なコンテンツベース方式によるフィッシング検知システムの実装を行った．HTMLソースからコンテンツの文字情報のみを抽出するため，正規表現によってHTMLタグをスペース文字に置換して除去している．キーワード抽出のための形態素解析器にはTreeTagger[7]を利用した．検索キーワードには，検査対象ページに出現した各名詞のTF-IDF上位 5 単語を選定した．TF-IDFの計算には，Yahoo!検索Web APIを用いた．IDFの計算には，サンプルドキュメント総数と，そのうちの当該単語を含むドキュメントの数が必要である．ここでは，サンプルドキュメントをYahoo!検索Web APIサービスが保持している総ウェブページ数，当該単語を含むドキュメント数を当該単語で検索した時のヒット数とした．

また，日本語のウェブページに適用可能なシステムの実装を行った．基本的な処理は先に述べたものと同様である．キーワード抽出には，日本語で利用可能な形態素解析器であるMeCab[8]を用いた．

5. 予備実験

5.1 概要

予備実験では，本実験に向けて本システムの調整を行うために実施するものである．ここでは，27 件のフィッシングサイト実例データについて本システムを適用し，その結果を分析した．

5.2 実験方法

実験で用いるサンプルデータは，過去のフィッシングサイトのデータのコピーをローカルに保存したものである．フィッシングサイトは存続期間が短いという性質があるため，今回のサンプルデータのフィッシングサイトは既にウェブ上から消滅していると考えられる．すでに存在しない URL を検索することは不可能であるため，フィッシング検知率は通常より高くなることが予想される．この点において，既存研究と比較評価することは不可能である．

そこで，今回の実験では，新たに次の点に着目する．

- **正規サイト導出率** - フィッシングサイトを検査し，検査対象ページから抽出したキーワードによる検索結果中に，検査対象ページの模倣元である正規サイトが含まれていた率

模倣元である正規サイトが検索できた場合は，特徴的な単語が正しく抽出できていると言える．この結果から，単語抽出の精度を評価することができる．

Yue らの実験結果より，フィッシング検知率，正規サイト誤検知率のいずれについて最も検知性能の高い方法は「Basic TF-IDF + domain +ZMP」であることが示されている．しかし，今回の実験では，正規サイト導出率を測定する上で，フィッシングサイトのドメイン名を加えることは，評価の妨げとなる．そこで，予備実験では，Yue らによる実験方法「Basic TF-IDF + ZMP」に倣った方法で行った．

5.3 実験結果と考察

実験結果を図 3 に示す．実験に使用したサンプルデータ 27 件のうち，26 件をフィッシングサイトと判断し，1 件を正規サイトと誤判断した．フィッシングサイトと判定した 26 件のうち，正規サイトのドメインが検索結果に含まれていたのは，11 件であった．これは，本システムの想定した通りの正しい動作を経て，正しい判断をしている場合である．残りの 15 件については，正規サイトが導出できていなかった．その要因について，分析を行った．

Vol.2010-DPS-142 No.44
Vol.2010-CSEC-48 No.44
2010/3/5

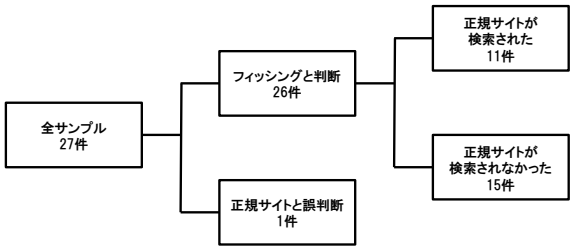

図 3 実験結果

(i) サブドメインを考慮することで解決するケース

本システムでは，模倣元の正規サイトと完全一致するドメインが検索結果に含まれていた場合に，正規サイトが検索されたと判断している．この条件を，サブドメインの一致とすることで，正規サイトが検索されたと判断できるケースが含まれていた．

(ii) HTML のパースの失敗

HTML のパースに失敗し，本来では出現しない語句が検索キーワードに選定された．本システムでは，正規表現によって HTML タグを半角スペースに置換して除去しているため，タグの位置によって単語が分割され，不自然な文字列が発生してしまう．

例：「Forgot」⇒「F」，「orgot」

また，タグ中の文字列に「<」「>」が使われており，HTML の規約に違反した記述方法をしていたために，パースに失敗した．

例：<form name=”Form1” method=”post”
action”Default.aspx” id=”Form1”
onsubmit=”if(email_user.valuse!=’<%=txtEmail%>’)return
isEmail(email_user);”>

(iii) ダイアクリティカルマークを含む文字による形態素解析の失敗

「español」「français」といった単語には，ダイアクリティカルマークを含む特殊な文字が出現している．本システムで利用している形態素解析器はこのような文字に対応しておらず，未対応の文字を境に単語が分割されてしまう．このようにしてできた単語が，検索キーワードに選定された．

例：「español」⇒「espa」，「ol」

(iv) 文字情報の少ないウェブページ

検査対象ページが，URL 転送のためのページや，frame タグによるフレーム操作を行うためのものである場合，本来のコンテンツが含まれておらず，特徴的なキーワードが抽出できない場合が多い．また，画像が多く文字情報の少ないウェブページについても同様のことが言える．本システムでは，こうしたウェブページについても，通常通りに形態素解析を試みたため，検索キーワードが抽出できなかった．

(v) 正規サイトと異なる特徴を有するウェブページ

検査対象ページ中に，正規サイトには見られない不自然な工夫が見られた．

例：「YAHOO」⇒「YAH00」（O を数字のゼロに置き換えている）

この方法が用いられたウェブページから抽出された検索キーワードには，模倣元の正規サイトにはない特徴が含まれる．

また，検査対象ページが無料ホスティングサービスを利用している場合，自動的に付加される広告中の語句が検索キーワードに選定された．

これらの語句は，正規サイト中には現れない．そのため，正規サイトが検索されなかった理由は明らかであり，キーワード抽出の精度向上によって解決することは不可能である．

(vi) 特徴的な語句が検索キーワードに選定されないケース

形態素解析は成功しているが，特徴的な語句が抽出できていないため，検索結果が雑然としており，フィッシングサイト，正規サイトのいずれの URL も含んでいなかった．

また，正規サイトと誤判断した１件については，(v)で述べた方法によって正規サイトと異なる特徴を有していたもので，かつフィッシングサイトが長期に渡り存在していたために，フィッシングサイトが検索されてしまった．そのため，(v)への対応策を行うことで，解決するものと考えられる．

5.4 改良

前節で述べたそれぞれのケースの対応策を検討した．

(A) サブドメインの一致を正規サイトとみなす

(i)への対応策となる．ドメイン比較時にサブドメインが一致した場合においても正規サイトと判断する．

(B) テキストブラウザ Lynx による HTML パース

(ii)への対応策となる．テキストブラウザである Lynx[9]のレンダリング結果を用いることで，ウェブページから文字情報のみを正しく取り出すことが可能となることが期待される．

(C) ダイアクリティカルマークを含む文字の除外

(iii)への対応策となる．ダイアクリティカルマークを含む文字を除外する．除外の方法は次のパターンを試みる．

Vol.2010-DPS-142 No.44
Vol.2010-CSEC-48 No.44
2010/3/5

- **単語除去モード** - ダイアクリティカルマークを含む単語を除去する．
- **文字置換モード** - ダイアクリティカルマークを含む文字を代替可能な文字で置換する．

文字置換モードの例：「español」⇒「espanol」

(D) **URL 転送やフレームページへの適用**

(iv)で述べたうちの，URL 転送およびフレームページへの対応策となる．これらのウェブページについては，転送後およびフレーム内に表示されているウェブページに対して本処理を行う．

(E) **ウェブページのタイトルを検索キーワードに加える**

(iv)で述べたうちの，画像が多く文字情報の少ないウェブページへの対応策となる．文字情報が少なく，抽出した単語が一定個数以下の場合において，ウェブページのタイトルを検索キーワードに加える．

6. 本評価

6.1 概要

本実験では，Yue らの実験方法のうち，フィッシング検知率の高い「Basic TF-IDF」「Basic TF-IDF + ZMP」「Basic TF-IDF + domain +ZMP」の３つの方法で実験を行った．

また，前章で述べた改良を施したうえで，英語・日本語からなる 843 件のフィッシングサイト実例データについて実験を行った．言語によって形態素解析器を選択する必要があるため，前処理として言語判定を行う処理を加えた．言語判定器には Lingua::LanguageGuesser[10]を用いた．

6.2 実験方法

次の方法による実験を行い，フィッシング検知率を測定する．

- Basic TF-IDF
- Basic TF-IDF + ZMP
- Basic TF-IDF + domain +ZMP

また，「Basic TF-IDF + ZMP」の結果については，正規サイト導出率の測定を行う．

前章で述べた改良点のうち，(B)テキストブラウザ Lynx による HTML パース，(C)ダイアクリティカルマークを含む文字の除外，については，これらの手法を使用するかどうかが選択可能である．そのため，表 1 に示す６通りの実験を行い，それぞれの結果を比較分析する．

表 1 実験モード

ダイアクリティカルマーク除外 ＼ タグ除去	正規表現	Lynx
行わない	①	④
(A)除去モード	②	⑤
(B)置換モード	③	⑥

表 2 フィッシング検知率

	フィッシングサイト	正規サイト	不明	フィッシング検知率
Basic TF-IDF	826	0	17	98%
Basic TF-IDF + ZMP	843	0	0	100%
Basic TF-IDF + domain +ZMP	843	0	0	100%

6.3 フィッシング検知率

実験結果を表 2 に示す．これは，実験モード①による実験結果である．全ての実験方法において，正規サイトと判断されたものは 0 件であった．また，ZMP を用いたいずれの方法において，フィッシングサイトと判断されたものは 0 件であった．

6.4 正規サイト導出率

実験結果を図 4 に示す．これは，実験モード①による結果である．「Basic TF-IDF + ZMP」による実験方法において，正規サイトが導出できた数は 698 件であった．

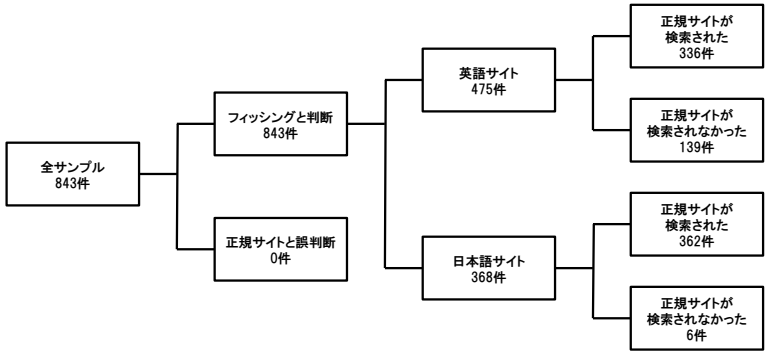

図 4 実験結果

6.4.1 モード別の比較評価

模倣元の正規サイトを検索することができた数について，各モード別に比較を行った．HTML パース方法によって実験結果が大きく異なるため，それらを分類した上で結果を示す．

(1) 正規表現を用いた HTML パース（①，②，③）での結果

正規表現によって HTML タグを除去するモードの中で，最も多くの正規サイトが導出されたものは②除去モードであり，その数は 843 件中 699 件（82.9%）であった．また，その他のモード（①，③）で正規サイトが導出されたものは全て，②での結果に含まれていた（図 5）．

Vol.2010-DPS-142 No.44
Vol.2010-CSEC-48 No.44
2010/3/5

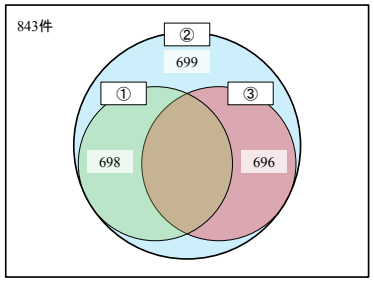

図 5 正規表現を用いたモードで正規サイトを検索できた数

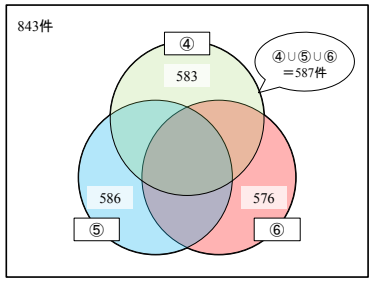

図 6 Lynx を用いたモードで正規サイトを検索できた数

(2) Lynx を用いた HTML パース（④，⑤，⑥）での結果

Lynx を用いたモードの中で，最も多くの正規サイトが導出されたものは⑤除去モードであり，その数は 843 件中 586 件であった．また，その他のモード（④，⑥）と和集合をとることで，結果は 1 件増の 587 件（69.6%）となった（図 6）．

これらの結果より，全てのモードの中で，最も多くの正規サイトが導出されたものは，②正規表現を用いた HTML パースおよびダイアクリティカルマークを含む単語を除去するモードであった．

(3) HTML パース方法による差異の考察

正規表現を用いたモードと Lynx を用いたモードの間で，実験結果が大きく異なった．これは，Lynx が html ソースに記述されているフォームや画像やその他の要素を，状況に応じてあらゆる文字列（[INLINE]や[ENBED]，[BUTTON]等）に置換して出力しており，これらの文字列がキーワードに選定されてしまったためである．そのため，Lynx を用いる際は，置換後の文字列を改めて除去する必要があった．

一方で，正規表現を用いたモードによって正規サイトが検索された 699 件と，Lynx を用いたモードによって正規サイトが検索された 587 件の和集合をとることで，正規サイトが検索された件数は 36 件増の 735 件（4.3%増の 87.2%）となった．これより，Lynx が行うレンダリング方法を参考にして，正規表現による HTML パース方法を調整することで，特徴的な単語の抽出精度を高めることが可能であると考えられる．

6.4.2 ダイアクリティカルマーク除外による成果

正規表現を用いたモード及び Lynx を用いたモードのどちらについても，より多くの正規サイトが検索されたものは，除去モードであった．また，置換モードは，ダイアクリティカルマークを除外しない結果より，結果が劣ることが分かった．これは，汎用性のない複雑な単語が英文字に置換されることで，検索結果に悪影響を及ぼすからであると考えられる．一例として，人名「Börzsönyi」を「Borzsonyi」と置換することで，検索件数が大幅に減っている事例があった．

6.4.3 その他の手法による成果

(1) タイトルキーワード

今回の実験では，タイトルキーワード手法の適用条件を，抽出された単語数が 6 件以下の場合とした．

本手法が適用されたものは 843 件中 7 件であり，いずれも画像によって構成されたウェブページであった．本手法によって，正規サイトが検索されたものは，7 件中 6 件であった．また，本手法を使用しない場合に，正規サイトが検索されたものは，7 件中 1 件であった．本手法を用いたことによって，結果が悪化したものはなかった．

(2) URL 転送及び frame を利用したウェブページへの適用

URL 転送及び frame を利用したウェブページを検出することにより，そのままシステムを適用したことで単語が抽出されないという事態を回避することに成功している．

6.4.4 正規サイトが導出されなかったケース

- **正規サイトと異なる特徴を有するウェブページ**

5.3 節(v)で述べたものと同様である．これらのウェブページは，正規サイトにはない特徴が見られる．そのため，特徴的な単語の抽出精度を上げることによって，これらのフィッシングサイトから正規サイトを導出することは不可能である．

このようなフィッシングサイトがウェブ上に長期的に存在した場合は，注意が必要である．なぜなら，長期的にウェブ上に存在することで検索エンジンの評価が高くなり，検索結果の中にフィッシングサイトが含まれる可能性が高まるからである．

- **検索エンジンに登録されないページ**

オンラインバンキングのログインページなどは，robots.txt や meta タグを用いることで，検索エンジンに登録されないようにされているケースがある．よって，このようなウェブページを検査した場合，抽出した語句でウェブ検索を行っても，そのページを検索することはできない．これについては，中山らが対応策を検討している[＝]．

- **特徴的な語句が検索キーワードに選定されないケース**

5.3 節(vi)で述べたものと同様である．形態素解析が失敗していないにも関わらず，企業名や商品名等の特徴的な語句が抽出されていないケースである．

このケースは，企業名が一般的な語句の組み合わせから成るウェブサイトで多く見られた．例えば，「Bank of America」（米国）は，企業名が「Bank」「America」といった非常に一般的な単語から成っている．そのため，それらが特徴的な単語として選定されないことで，正規サイトが導出されなかった．同様のケースに当てはまる企業のウェブサイトとして，「Alliance & Leicester」（英国）等がある．

その他にも，ウェブページで用いられているシステムのヴァージョン番号の表記や，ID やパスワードの入力例として出現する文字列等の，汎用的でない単語が特徴語として選定されるケースがあった．これにより，本来選定されることが望ましい単語が検索キーワードに選定されない．こうした問題への対応は，今後の課題である．

Vol.2010-DPS-142 No.44
Vol.2010-CSEC-48 No.44
2010/3/5

7. 結論

コンテンツベース方式は，データベースが不要で，かつ即時性の高いフィッシング検知方式として注目されている．しかし，その検知性能については小規模な評価しか行われておらず，大規模な実例データを利用した評価が行われていなかった．また，日本語に対応したシステムの実装が行われておらず，他言語での有効性が評価されていなかった．そこで本研究では，日本語に対応するシステムを実装し，大量の日英のフィッシング実例データによる適用実験を行った．

コンテンツベース方式は，文字情報が少ないウェブページの検査が行えない．文字情報が少ないウェブページとは，(1)URL 転送のためのウェブページ，(2)frame タグによるフレーム定義を行うウェブページ，(3)画像が多く文字情報の少ないウェブページなどがある．そこで，(1)，(2)については，転送後及びフレーム内に表示されている URL を HTML ソース中から見つけ出し，それらのウェブページに本処理を適用することで対応した．(3)については，ページのタイトルを検索キーワードとすることで，本処理を行う方式を実装した．これより，文字情報の少ないウェブページへの対策が有効に機能していることを確認した．

また，HTML パース方法および特殊文字の扱い方が検知性能に与える影響について調査するために，(1)正規表現により HTML タグを半角スペースに置換する方法，(2)テキストブラウザ Lynx によるレンダリングを利用する方法，を実装評価した．これより，テキストブラウザ Lynx によるパース方法を参考にすることで，特徴的な単語抽出の精度が向上できる見通しを得た．

実験結果から，英語，日本語のいずれのフィッシングサイトについても，高い検知性能が示された．これにより，コンテンツベース方式の有効性を確認した．また，正規サイトが導出されなかった原因の分析を行ったところ，(1)フィッシングサイトが正規サイトと異なる特徴を有しているケース，(2)検索エンジンに登録されないページ，(3)特徴的な語句が検索キーワードに選定されないケース，であることが分かった．(3)については，正規サイトをフィッシングサイトであると判断する原因と成りうる．そのため，検査対象ページとそこから実際に選定された単語を比較し分析することで，単語の抽出精度を高める方法を検討する必要がある．

謝辞 本研究は、フィッシング対策協議会（事務局 : JPCERT コーディネーションセンター）からの受託研究により行われたものである．

参考文献

1 Gartner Survey Shows Phishing Attacks Escalated in 2007; More than $3 Billion Lost to These Attacks, http://www.gartner.com/it/page.jsp?id=565125（2010 年 1 月確認）
2 Yue Zhang, Jason Hong, Lorrie Cranor, “CANTINA: A Content-Based Approach to Detecting Phishing Web Sites”, WWW2007, (2007).
3 中山 心太，吉浦 裕, “模倣コンテンツの特性に基づくフィッシング検知方式”, 2007-CSEC-38, Vol.2007, No71, pp387-392, (2007).
4 柴田 賢介, 荒川 陽助, 塩野 入理, 金井,“Web サイトからの企業名抽出によるフィッシング対策手法の提案”, IPSJ SIG Notes Vol.2006, No.96 pp.17-22(2006).
5 RBL.JP, http://www.rbl.jp/
6 Phishing Activity Trends Report for the Month of January, 2008, http://www.antiphishing.org/reports/apwg_report_jan_2008.pdf（2010 年 1 月確認）
7 TreeTagger, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
8 MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/
9 Lynx for Win32 (by patakuti): Project Home Page, http://lynx-win32-pata.sourceforge.jp/
10 Lingua::LanguageGuesser - 言語判定器, http://gensen.dl.itc.u-tokyo.ac.jp/LanguageGuesser/LanguageGuesser_ja.html