コンテンツベースフィッシング検知手法の実用化に向けた評価と改良

フィッシング対策協議会

2011 年 9 月 28 日

概要

コンテンツベースフィッシング検知方式は，検査対象ページのコンテンツからキーワードを抽出して，なりすまし元の正規サイトを検索し，対象ページと比較することでフィッシングを検知する．本研究では，コンテンツベース方式の実用化に向けて，正規サイト，携帯向けフィッシングサイト，JPCERT/CC より新たに提供を受けた PC 向けフィッシングサイトを用いた評価実験を行った．

正規サイト 100 件のうち 9 件についてフィッシングと誤検知した．これらの原因を分析した結果，HTML からのコンテンツ抽出手法の改良，N グラム法を用いたキーワード抽出の高度化，検査対象ページだけでなくその周辺ページを考慮するキーワード抽出の高度化（ドメインキーワード手法）によって解決可能であることが分かった．

17 件の携帯向けフィッシングサイトを用いた評価では，17 件全てに対して正しくフィッシング検知ができた．一方，模倣元である正規サイトを検索できたのは 1 件のみであり，16 件については検索できなかった．16 件のうち 7 件については，正規サイトの誤検知にはつながらないが，残りの 9 件についはて正規サイトの誤検知につながる可能性がある．その対策として，コンテンツの少ない携帯サイトからブランド名を正しく特定する，キーワード抽出の高度化が必要である．

最新の PC 向けフィッシングサイトへの適用評価を行う前に，正規サイト誤検知率の低減と正規サイト導出率の向上の，最も有効な方法と考えられるドメインキーワード手法を実装し，現在のコンテンツベース手法に組み込んだ．

新たに提供を受けた PC 向けフィッシングサイトへの適用評価では，154 件のフィッシングサイトを評価し，全て正しくフィッシング検知できた．一方，模倣元の正規サイトを正しく導出できたのは 121 件であった．正規サイトを導出できなかった 33 件のうち 32 件については，正規サイトの誤検知につながらないが，残りの 1 件については，キーワード抽出の高精度化による対策が必要である．

目次

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 年ではフィッシング攻撃が国内で過去最多数を記録した[2].

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[3]，中山ら[4]によって提案されているコンテンツベース方式は，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．この方式は，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．しかし，大規模な実例データを用いた実験を行っていないため，実用性が明らかではない.

前年度の受託研究では，JPCERT/CC の保有する 843 件のフィッシング実例データを用いてコンテンツベース方式の評価に対して大規模な実験を行った．その結果，フィッシング検知率が 100%であるが，正規サイトをフィッシングサイトと誤検知する可能性が明らかになった．また，この誤検知を防止する方式として，中山らが提唱するドメインキーワード手法の有効性が明らかになった.

そこで本稿では，前年度の評価に続いて，以下の 3 点の評価と改良を行った.

- 正規サイトへの適用評価
  - 100 件の正規サイトに適用し，正規サイト誤検知率を評価する.
- 携帯向けフィッシングサイトへの適用評価
  - 近年増加している携帯向けフィッシングサイトに適用し，フィッシング検知率を評価する．また，模倣元である正規サイトを正しく特定できた率(正規サイト導出率).
- 新たに提供を受けた PC 向けフィッシングサイトへの適用評価
  - JPCERT/CC より新たに提供を受けた PC 向けフィッシングサイトについて，フィッシング検知率と正規サイト導出率を評価する.

2. コンテンツベース方式

コンテンツベース方式とは，フィッシングサイトが正規サイトの模倣をしているという特性を利用したフィッシング検知システムである．

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）．

(1) 検査対象ページ内の各単語について特徴度を算出する．
(2) 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．
(3) 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．

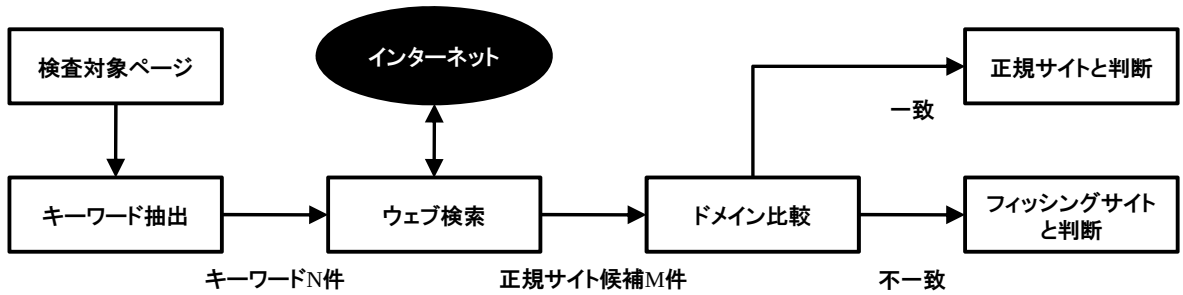

図 1 コンテンツベース方式

なお，上記（1）では，標準的な特徴度算出法である TF-IDF を用いる．コンテンツベース方式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 [1]，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．

2.1 評価方法

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，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

正規サイト導出率

フィッシングサイトを検査する際に模倣元である正規サイトを正しく特定できる率．これは，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とする．

3. 前年度の評価

日英両言語のコンテンツベースシステムを実装し，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，以下の結果を得た．

フィッシングサイト検知率については，843 件全てについて正しいフィッシング判定が得られ，検知率 100%であった．正規サイト導出率については，705 件のなりすまし元である正規サイトの検索に成功した．残りの 138 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．その結果，正規サイトでも同じ理由で誤検知をする可能性のあるものが 57 件であった．そのうちの 31 件が，中山らが提案したドメインキーワード手法において対応できる可能性があった．

なお，ドメインキーワード手法とは，検査対象ページから同一ドメインのリンクを辿り，そのリンクを検査対象ページとして検査することで，ブランド名等の特徴的な語句をキーワードとして選定する手法である．

4. 正規サイトへの適用評価

4.1 目的

正規サイトへの誤検知を低減するために ECHCO システムの改良点の発見することを目的とする.

4.2 実験方法

吉浦研究室が開発した ECHCO システムを, JPCERT/CC の保有する正規サイトの実例データ 100 件に適用し, その結果を分析する. また, 今回の実験では, 前年度の実験での 6 通りのモードの中で 1 番よかった②のモードによって実験を行った（表 1).

表 1 実験モード

タグ除去 / ダイアクリティカルマーク除外	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

4.3 実験結果

4.3.1 実験適用データの分類

100 件の実例データを分類したものを図 2 に示す.

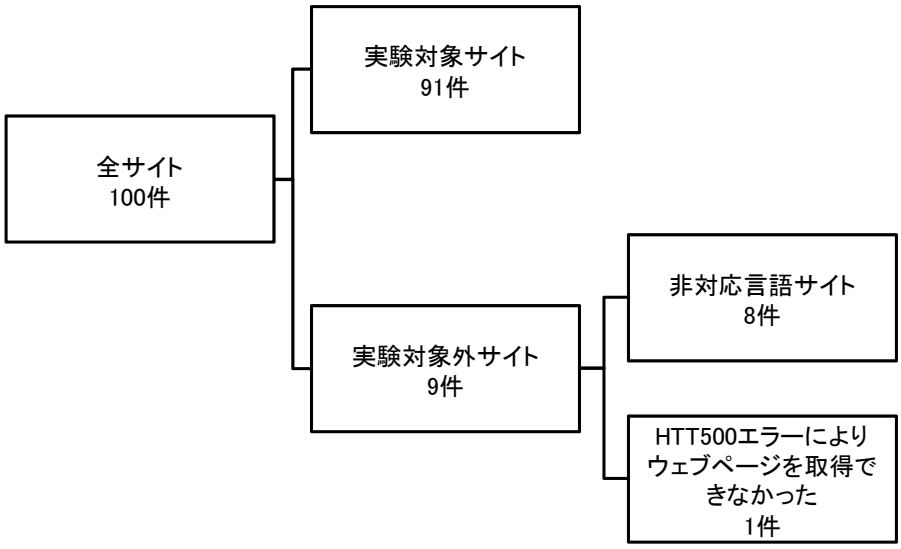

図 2 実験適用データの分類

日本語，英語以外の非対応言語サイトが 8 件あった．これらは対象外とした．なお，この判定は，言語判定器 Lingua::LanguageGuessor が自動的に行なった．8 件の内訳としては，スペイン語が 3 件，フランス語が 1 件，イタリア語が 4 件であった.

また，HTTP500 エラーによりウェブページが取得できなかったものが 1 件あった．これは，Web サイトがメンテナンス中か，Web サイトのプログラムに問題があるため，URL からウェブサイトのソースを持ってくることが出来なかった．そのため，このサイトは実験の対象としなかった.

4.4 検知結果

検査対象データ 91 件中，82 件を正規サイトであると判定し，残りの 9 件をフィッシングサイトであると誤判定した．検知結果の全体図を図 3 に示す．

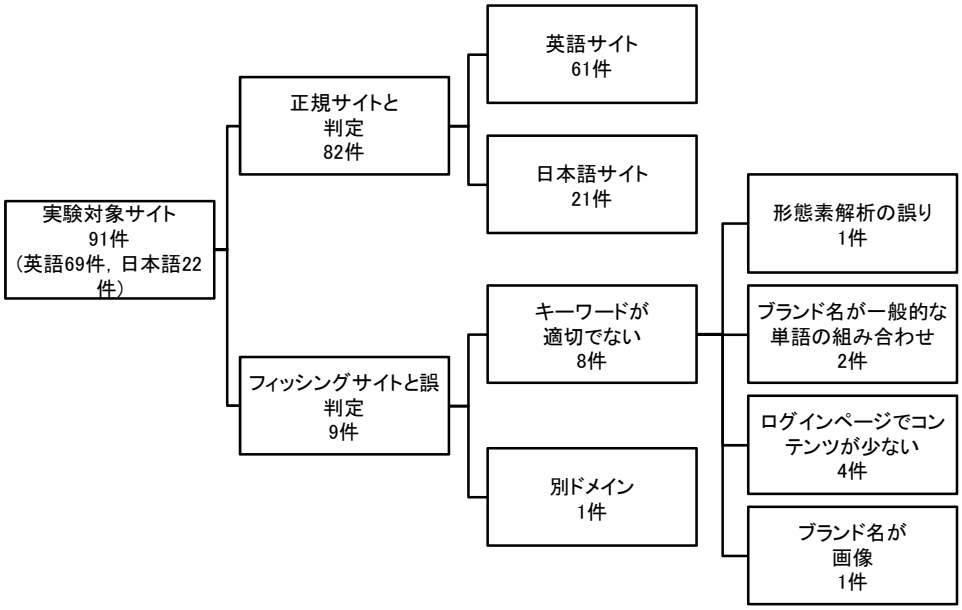

図 3 検知結果の全体図

以降は，個々のケースについて，詳細な分析を行ったものである．

4.5 分析

4.5.1 正規サイトと判定したケース

英語サイトについては 69 件のうち 61 件(88%)を正規サイトと正しく判定し，日本語サイトについては 22 件のうち 21 件(95%)を正規サイトと正しく判定した．

4.5.2 フィッシングサイトと誤判定したケース

フィッシングサイトと判定した 9 件に関して表 2 にまとめる．

表2 実験結果のまとめ

ブランド名	URL	概要	ECHCOシステムが実際に見つけたキーワード	人間が工夫し，検索に成功したときのキーワード	考察
ICICI Bank	https://infinity.icicibank.co.in/BANKAWAY?Action.RetUser.Init.001=Y&AppSignonBankId=ICI&AppType=corporate&abrdPrf=N	"infinity.icicibank.co.in"を一つの単語として形態素解析TreeTaggarをしてしまった.	"banking" "infinity.icicibank.co.in" "keyboard" "internet" "security"	"banking" "icici" "infinity" "password" "security"	形態素解析器であるTreeTaggerが"infinity.icicibank.co.in"を一つの単語として認識してしまった.
Old Point National Bank	https://portal1.secure-banking.com/60326/PassmarkSignIn.faces	ブランド名が一般的な単語の組み合わせであった.	"demo" "banking" "information" "sign" "email"	"demo "banking" "information" "sign" "old"	"Old" と"Point"と "National"と "Bank"がいずれも一般的な単語であり，キーワードに選ばれなかった.
Islamic Bank of Britain	https://online.islamic-bank.com/online/aspscripts/Logon.asp	ブランド名が一般的な単語の組み合わせであった.	"security" "banking" "emails" "details" "circulation"	"islamic" "banking" "britain" "log" "islamic"	"Islamic"と"Bank"と"Britain"がいずれも一般的な単語であり，キーワードに選ばれなかった.
ソースネクスト	https://www.sourcenext.com/users/action/mypage_chk3	ログインページ共通の単語"パスワード" "入力"等を拾ったため，ブランド名(ソースネクスト)が外れてしまった.	"ユーザー" "パスワード" "入力" "情報" "個人"	キーワードの6番目である"SOURCENEXT"が入っていれば検索可能	ログイン共通の"パスワード" "入力" などがキーワードに含まれたため，"SOURCENEXT"が含まれなくなった.
RBC Royal Bank	https://www1.royalbank.com/cgi-bin/rbaccess/rbunxcgi?F6=1&F7=IB&F21=IB&F22=IB&REQUEST=ClientSignin&LANGUAGE=ENGLISH	ログインページ共通の単語をひろったため，ブランド名がキーワードから外れてしまった.	"banking" "online" "rbc" "security" "enrol"	キーワードの6番目である"royal"が入っていれば検索可能	ログインページに共通の"banking" "online"などがキーワードに含まれたため，"royal"が含まれなくなった.
cahoot	https://ibank.cahoot.com/servlet/com.aquarius.security.authentication.servlet.LoginEntryServlet	ログインページ共通の単語が多く，キーワードの特徴が薄まった.	"cahoot" "aer" "reconfirm" "rate" "interest"	"cahoot"のみ，または"cahoot"と"aer"のみであれば検索可能	ログインページ共通の"reconfirm" "rate" "interest"の単語がキーワードに含まれたため検索できなかった.
Desjardins	https://accesd.desjardins.com/en/accesd/	ログインページ共通の単語が多く，キーワードの特徴が薄まった.	"security" "desjardins" "fraud" "demonstration" "browsers"	"security" と"desjardins"を用いれば検索できる	ログインページ共通の"fraud" "demonstration" "browsers"単語がキーワードに含まれたため検索できなかった.
Dade County Federal Credit Union	https://mfaweb.dcfcu.org/auth/Authorize?fiid=1	ブランド名が画像になっていた.	"authentication" "mfa" "dade" "enroll" "demonstration"	人間でも検索できない	ブランド名が画像になっているために，ブランド名をキーワードとすることが出来なかった.
Sabine State Bank	http://k2.secure-banking.com/1598.cfm	ブランド名(Sabine State Bank)をキーワードにしても，正規サイトが検索できなかった.	"sabine" "banking" "fdic" "bank" "state"	yahooAPIではできない．googleAPIでは"Sabine State Bank"で出現	検査対象のログインページとSabine State Bankのホームページが別ドメインであった.

以下原因毎にくわしく分析する

4.5.2.1 キーワードが適切でないケース

キーワードが特徴的でなかったり，適切でないために検索結果に正規サイトが出現しなかった．以下に詳細を述べる．

（1）形態素解析の失敗

ICICI Bank が該当した．

キーワードの一つに URL のドメイン名(infinity.icicibank.co.in)がそのまま入っていたため，このキーワードを含む 5 件で検索をしても正規サイトは出現しなかった．

（2）ブランド名が一般的な単語で構成

Old Point National Bank と Islamic Bank of Britain の 2 件が該当した．

いずれのブランドも"old"，"point"，" Islamic"，"Britain"等の一般的な単語で構成されていた．これらの単語は TF-IDF による特徴度が低かったのでキーワードに選ばれず，その結果正規サイトを検索できなかった．Old Point National Bank の場合，キーワードに手動で"old"を入れたら正規サイトが検索できた（ECHCO システムの選定した 5 つのキーワードの 5 番目"email"を"old"で差し替えた上で検索エンジンを起動した）．同様に，Islamic Bank of Britain の場合も"islamic"をキーワードに入れたら，正規サイトが検索出来た．

（3）ログインページでコンテンツが少ない

フィッシングサイトと判定した 9 件のうち，4 件が該当した．

ログインページの共通の単語として，"ログイン"，"入力"等がある．しかし，これらは検査対象ページの特徴的な単語ではないため，正規サイトの検索に寄与しない．一方，これらの単語がキーワードに入ることで，本来拾うべきキーワードが拾われなくなったしまった．その結果，検索結果に正規サイトが出現しなかった．

具体例として，「ソースネクスト」では，キーワードの 2 番目と 3 番目に"ログイン"，"入力"が入ったために"SOURCENEXT"が 6 番目に落ち，キーワードにならなかった．

同様に「RBC Royal Bank」では，"banking"と"online"が入ったために"royal"が 6 番目に落ち，キーワードにならなかった．

また，ログインページに共通の単語をキーワードに含んだため，キーワードの特徴がなくなり，正規サイトの検索を失敗したケースがあった．

「cahoot」は"cahoot"と"aer"だけで検索すれば正規サイトが出現するのに，これらに加えて，"reconfirm"，"rate"，"interest"という当たり前の単語がキーワードに加わってしまったため正規サイトが検索できなかった．

同様に「Desjardins」では，"security"と"desjardins"だけで検索すれば正規サイトが出現するのに，"fraud"，"demonstration"，"browsers"が入ってしまったため正規サイトが検索できなかった．

なお，「ソースネクスト」の場合ロボットタグが入っているため(<META name="robots"

content="noindex,nofollow">)，このログインページが検索結果に出現することはない．しかし，適切なキーワード("SOURCENEXT" "ユーザー"パスワード" "入力" "情報")を用いて検索すると，「ソースネクスト」社のトップページが現れ，ドメインが同じであることから正規サイトと判定できる．

（4）ブランド名が画像であったケース

Dade County Federal Credit Union が該当する．

ウェブページのブランド名が画像であったために，キーワードにブランド名が取得できず，検索結果に正規サイトが出現しなかった．

4.5.2.2 検索対象ページとブランドのホームページが別ドメイン

Sabine State Bank が該当する．

検索対象ログインページと銀行本体が別ドメインであったため，本体のホームページは検索されたがドメインが一致せずフィッシングと判定された．そして，ログインページ自身は検索されなかった．なお，GoogleAPI を用いるとログインページが検索でき，正規サイトと判定する．

4.6 改良案

今回の実験でフィッシングサイトと判定した 9 件を正規サイトと判定するための改良案を以下に示す．

4.6.1 タグ除去時の正規表現の改良

今回の実験で出来なかった 9 件のうち ICICI Bank に対応できると思われる．
空白を含まない 2 つの文字列 A，B がピリオドでつながっているときに，そのピリオドを空白に変える．(A．B→A　B)

キーワードがドメイン名(infinity.icicibank.co.in)になっていたので，ピリオドを空白にできていれば，「infinity」「icicibank」というように分解できるので，正規表現を改良すればこの問題を解決できるはずである．

4.6.2 N グラム法の導入

Old Point National Bank や Islamic Bank of Britain のように，ブランド名が一般的な単語で構成されている場合，ブランド名中の個々の単語は特徴度が低くなり，キーワードとして選定されない．そのため，" Old Point National Bank"のような複数の単語のつながりを一つのキーワードとして選定する必要がある．そのような手法として，N グラム法の利用が考えられる．

N グラム法は，任意の長さ N の文字列を一つの単語とみなすテキスト分析の手法である．たとえば，"abc dbc dabc"という文字列に，3 グラム法を適用すると，"abc"，"bc "，"c d"，" db"，"dbc"，"bc "，"c d"，" da", "dab", "abc"が単語として認識され，"abc"，"bc "，"c d"

の出現頻度が各 2 回, " db", "dbc", " da", "dab"の出現頻度が各 1 回という結果が得られる.

"old point national bank"は 23 文字なので，23 グラム法を用いれば，一つの単語として認識され，キーワードとして使えるようになると考えられる．その結果，今回の実験で出来なかった 9 件のうち Old Point National Bank と Islamic Bank of Britain に対応できると考えられる.

4.6.3 ドメインキーワード手法の導入

今回の実験で出来なかった 9 件のうちソースネクスト，RBC Royal Bank，cahoot，Desjardins，Dade County Federal Credit Union の 4 件に対応できると思われる.

ドメインキーワード手法とは，中山が 2009 年に考案したもので，検査対象ページからリンクをたどって同じドメインの周辺のページを調べ，ドメイン全体の特徴的なキーワード得るという手法である．それによって，検索結果に正規サイト，またはそのドメインを含んだサイトが出現して，問題を解決できるはずである.

上記以外に，4.5.2.2 節で述べたように，Sabine State Bank の場合，検査対象ページが同銀行のログインページであり，同銀行のホームページは検索できたが，検査対象ページは検索できなかった．そして，ホームページとログインページのドメインが異なっていたため, フィッシングサイトと誤検知した. ここで, ECHCO システムが用いている Yahoo!API の代わりに Google API を手動で利用してみたところ，ログインページとホームページの両方が検索できた.

このことは，検索エンジンによって，インデックスやランキングアルゴリズムが異なるため，検索結果が異なり，一方の検索エンジンでは正検知，一方では誤検知になる可能性を示している. 検索結果が広いほうが誤検知の可能性が小さくので, 可能であれば, ECHCO システムにおいて，Yahoo!API と Google API を併用し，両方ともフィッシングと判定した場合のみフィッシングと最終判定することが望ましい.

5. 携帯向けフィッシングサイトへの適用評価

5.1 概要

携帯フィッシングサイトの特性は PC サイトとは大きく異なる. 以下に分析結果の概要を述べる.

1. 今回サンプルとしたフィッシングサイトの大半を占めるモバゲーポイントは，正規サイトの模倣ではないようである．我々の調べた限り，正規サイトであるモバゲータウンにはモバゲーポイントというページが見あたらない.
2. 携帯のサイトは PC サイトに比べてページが簡易なため，正規サイトを模倣しなくてもフィッシングサイトを簡単に作れることが上記の原因と思われる.
3. 検査対象のフィッシングサイトとそっくりの別サイトがある．これもフィッシングサイトのようである．同じ攻撃者が複数の URL でフィッシングサイトを立てているようである.
4. 現在 ECHCO システムが利用している Yahoo!API は携帯サイトの検索ができない．携帯サイトを検索可能な Google/m は，我々の調査した限り API を公開していないようである．そこで，ECHCO システムのうち正規サイト検索の部分について，Google/m を手動で利用した．それ以外に，Yahoo!API での自動検索も行なった.
5. フィッシングの検知は，Google/m でも Yahoo!API でも 100%できた.
6. 正規サイトの検索は Yahoo!API では殆どできなかった．これは，Yahoo!API が携帯サイトを検索できないためである．フィッシングサイト（モバゲーポイント）については Google/m を用いても正規サイトを検索できなかった．これは，フィッシングサイトが正規サイトに似ていないためと考えられる．判定対象が正規サイトの場合には，この状況にならないので，誤検知を示唆するものではない.
7. ECHCO システムが自動抽出したキーワードを調べると，ブランド名そのものは抽出できていなかった．たとえば，フィッシングサイト（モバゲーポイント）の場合，「モバゲーポイント」というキーワードを抽出していたが，「モバゲー」「モバゲータウン」は抽出していない.
8. ブランド名そのものをキーワードとすると，たとえ Yahoo!API を用いても，正規の PC サイトが検索される. 正規の PC サイトは正規の携帯サイトと同一ドメインにあるため，ドメインレベルでは，正規サイトの検索に成功する.
9. 以上から ECHCO システムの改良方針として，以下が導かれる.

   ・Google/m の API を利用できれば，フィッシング検知率 100%となる．また，正規サイト検索率も，フィッシングサイトが正規サイトの模倣でない場合を除くと 100%となる．フィッシングサイトが正規サイトの模倣でない場合は，誤検知の対象にならないので問題ない.

   ・キーワード抽出処理を高度化し，ブランド名そのものを抽出できれば，現在の Yahoo!API を用いても Google/m と同じ性能が得られる.

5.2 目的

携帯のフィッシングサイトに対する ECHCO システムの性能を明らかにする.

5.3 実験方法

5.3.1 概要

吉浦研究室が開発した ECHCO システムを，JPCERT/CC の保有する携帯サイトのフィッシング実例データ 17 件に適用し，その結果を分析する.

5.3.2 実験データ

全て日本語の携帯 SNS サイトへのなりすましである. なりすまし先の内訳は以下の通り.

mixi 1 件，モバゲータウン 12 件，ixen 2 件，Gree 2 件

5.3.3 評価項目

フィッシング検知率：フィッシングと正しく検知した率

正規サイト検索率：模倣元の正規サイトを正しく検索できた率

5.4 実験を進めるにあたっての問題点

ECHCO システムは Yahoo!API を用いて正規サイトを自動検索しているが，Yahoo!API では PC サイトのみ検索可能であり，携帯サイトは検索できない．携帯サイトに対応した Google の検索エンジン(以後 Google/m と呼ぶ)がある[5]．これは Google 検索の携帯版であり，PC から携帯サイトを検索し，見ることができる．しかし，我々の調べた範囲では Google/m の API はなかった.

そこで，ECHCO システムのうち検索の部分は Google/m を用いて手動で行った．また，Yahoo!API を用いた ECHCO システムでの自動検知も別途行った.

5.5 検知結果

5.5.1 概要

検査対象データ 17 件に全てについてフィッシングサイトであると正しく判定できた.

正規サイト検索率は以下のようであった.

- 正規サイト検索率（Yahoo 利用）: 17 件中 1 件（5.88%）
- 正規サイト検索率（Google/m 利用）: 17 件中 9 件（52.94%）

正規サイト検索率（Yahoo 利用）が低いのは，Yahoo から携帯サイトが検索できないためである．Googl/m による実験結果の全体図を図 4 に示す.

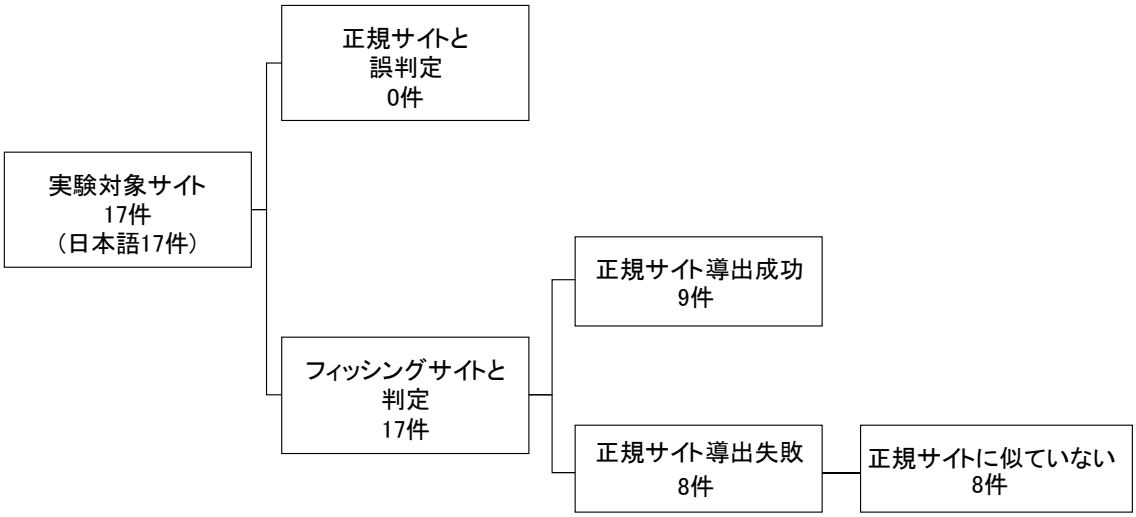

図 4 Google/m を用いた検知結果の全体図

5.5.2 **詳細**

詳細な分析結果を表 3 に示す．実験結果は，表 3 の 1 列目（種別）に示すように，A,B,C の 3 パターンに大別される．

A : 正規サイトが検索され，フィッシングサイトは検索されなかった．すなわち正常ケースである．

B : 正規サイトが検索されたが，検査対象とは別のフィッシングサイトらしきものが検索された．

C : 正規サイトもフィッシングサイトも検索されなかった．

表 3 のその他の列は下記を示している．

- ID：実験サンプルの番号
- ブランド：模倣元の企業名
- 検査対象 URL：フィッシングサイトの URL
- ECHCO システムが見つけたキーワード：TF-IDF 値の上位 5 件
- 特徴的なキーワードの出現順位：ブランド名やブランド名を含む語句の TF-IDF 値が何番目であったか.（ECHCO システムは TF-IDF 値 5 番目までをキーワードとして抽出する）
- 検査対象サイト：実験で用いたフィッシングサイトの検索結果．検査対象サイトが 50 位以内に出ていない場合にはフィッシングと正しく判定出来ている．
- 正規サイト：フィッシングサイトの模倣元のサイトの検索結果．正規サイトが 50 位以内に出ている場合には正規サイトの検索に成功している．
- 他のフィッシングサイト：実験で用いたフィッシングサイトに類似したサイトの検索結

果. 検査対象サイトではないが，フィッシングサイトらしいサイトが 50 位以内に出てくる場合がある．これは，もしそのサイトが検査対象サイトになった場合には，見逃す可能性があるということを意味している．

- 5 番目（一番下）のキーワードを外し，代わりに検査対象サイトのページタイトルを 1 番のキーワードに入れて検索した結果．このタイトルキーワード手法の有効性を評価している．1 行目はタイトル名，2 行目は正規サイトの検索結果，3 行目は他のフィッシングサイトの出現結果.

表3 実験結果のまとめ

種別	ID	ブランド	検査対象 URL	ECHCO システムが見つけたキーワード	特徴的なキーワードの出現順位	検査対象サイト	正規サイト	他のフィッシングサイト	タイトルキーワード入れて検索(5 位を外す)
A	17760	mixi	http://gctinc.net/mixi/	″登録″ ″メール″ ″利用″ ″下記″ ″mixi″	mixi5 位	100 件内に出現せず	1 番目に mixi のサイト出現	100 件内に出現せず	タイトルは″mixi″ 1 位に正規サイト出現
B	17763	モバゲーTOWN	http://951k.net/moba/	″登録″ ″利用″ ″お客様″ ″メール″ ″モバゲーポイント″	モバゲーポイント 5 位, モバゲー7 位	出現せず(75 件)	17 番目にモバゲータウンのサイト出現	6 番目に出現	タイトルは″モバゲーポイント″ 18 位に正規サイト出現 6 番目に他のフィッシング出現
C	17764	モバゲーTOWN	http://egajaty.net/	″登録″ ″メール″ ″利用″ ″換金″ ″送信″	モバゲーポイント 9 位, モバゲー22 位	100 件内に出現せず	100 件内に出現せず	100 件内に出現せず	タイトルは″モバゲーポイント″ 18 位に正規サイト出現 5 番目に他のフィッシングサイト出現
C	17765	モバゲーTOWN	http://dyedegd.net/	″登録″ ″メール″ ″利用″ ″換金″ ″送信″	モバゲーポイント 9 位, モバゲー22 位	100 件内に出現せず	100 件内に出現せず	100 件内に出現せず	タイトルは″モバゲーポイント″ 18 位に正規サイト出現 5 番目に他のフィッシング出現
C	17766	モバゲーTOWN	http://ebsjels.net/	″登録″ ″メール″ ″利用″ ″換金″ ″送信″	モバゲーポイント 9 位, モバゲー22 位	100 件内に出現せず	100 件内に出現せず	100 件内に出現せず	タイトルは″モバゲーポイント″ 18 位に正規サイト出現 5 番目に他のフィッシング出現
B	17767	モバゲーTOWN	http://lynnswannforgovernor.com/	″登録″ ″利用″ ″お客様″ ″メール″ ″モバゲーポイント″	モバゲーポイント 5 位, モバゲー7 位	出現せず(75 件)	17 番目にモバゲータウンのサイト出現	6 番目に出現	タイトルは″モバゲーポイント″ 18 位に正規サイト出現 6 番目に他のフィッシング出現
B	17973	モバゲーTOWN	http://axyhdue.net/navi/	″登録″ ″利用″ ″お客様″ ″メール″ ″モバゲーポイント″	モバゲーポイント 5 位, モバゲー7 位	出現せず(75 件)	17 番目にモバゲータウンのサイト出現	6 番目に出現	タイトルは″モバゲーポイント″ 18 位に正規サイト出現 6 番目に他のフィッシング出現

C	17980	モバゲータウン	http://xp-10.net/mg/	"登録" "メール" "お客様" "必須" "完了"	モバゲーポイント 13 位, モバゲー14 位	100 件内に出現せず	100 件内に出現せず	100 件内に出現せず	タイトルは"迷惑メールを送らないようにする方法" 検索結果 0 件で出現せず
A	18054	ixen	http://beefeaterlondonart.com/	"登録" "ixen" "無料" "利用" "メール"	ixen2 位	100 件内に出現せず	1 番目に ixen の正規サイト出現	100 件内に出現せず	タイトルは"ixen" 1 位に正規サイト出現 他のフィッシングサイトは出現せず
B	18120	モバゲータウン	http://ilwu24fcu.com/mobagee/	"登録" "利用" "お客様" "メール" "モバゲーポイント"	モバゲーポイント 5 位, モバゲー7 位	出現せず(75 件)	17 番目にモバゲータウンのサイト出現	6 番目に出現	タイトルは"モバゲーポイント" 18 位に正規サイト出現 6 番目に他のフィッシング出現
A	18121	ixen	http://legowebmark.net/	"登録" "ixen" "無料" "利用" "メール"	ixen2 位	100 件内に出現せず	1 番目に ixen のサイト出現	100 件内に出現せず	タイトルは"ixen" 1 位に正規サイト出現 他のフィッシングサイトは出現せず
B	18266	モバゲータウン	http://ilwu24fcu.com/1001/mobagee/	"登録" "利用" "お客様" "メール" "モバゲーポイント"	モバゲーポイント 5 位, モバゲー7 位	出現せず(75 件)	17 番目にモバゲータウンのサイト出現	6 番目に出現	タイトルは"モバゲーポイント" 18 位に正規サイト出現 6 番目に他のフィッシング出現
C	18490	モバゲータウン	http://ds1bn7dfnd.net/	"登録" "メール" "利用" "換金" "送信"	モバゲーポイント 10 位, モバゲー24 位	100 件内に出現せず	100 件内に出現せず	100 件内に出現せず	タイトルは"モバゲーポイント" 18 位に正規サイト出現 5 番目に他のフィッシング出現
C	18508	モバゲータウン	http://ds1bn7dfnd.net	"登録" "メール" "利用" "換金" "送信"	モバゲーポイント 10 位, モバゲー24 位	100 件内に出現せず	100 件内に出現せず	100 件内に出現せず	タイトルは"モバゲーポイント" 18 位に正規サイト出現 5 番目に他のフィッシング出現

A	18651	Gree	http://mail-gree.jp/	"登録" "gree" "場合" "確認" "必須"	gree2 位	100 件内に出現せず	3番目にGREEのサイトのサブドメインが出現	100 件内に出現せず	タイトルは"mail-gree.jp" 1 位に正規サイトとサブドメイン合致するサイトが出現
A	18676	モバゲーTOWN	http://showtrip.net/mg/	"確認" "必須" "メール" "お客様" "登録"	モバゲーポイント 12 位, モバゲー18 位	100 件内に出現せず	100 件内に出現せず	100 件内に出現せず	タイトルは"モバゲーポイント" 全 12 件で出現せず
C	19721	Gree	http://mobile-start.net/g98/	"登録" "会員" "メール" "利用" "送信"	gree8 位	100 件内に出現せず	100 件内に出現せず	100 件内に出現せず	タイトルは"tv でお馴染み gree からの特別キャンペーンのお知らせ！" 1 位に正規サイトとサブドメイン合致するサイトが出現

マーク A がついているものは，Google/m で正規サイトの検索に成功し，かつフィッシングらしき別サイトの検索をしなかったものである．これは正常ケースである．

マーク B がついているものは，Google/m で正規サイトの検索に成功したが，フィッシングサイトらしき別サイトを検索したものである．17763（表 3 の ID 欄参照）などでは，キーワードに「モバゲーポイント」が入っていたために正規サイトが検索されたが，上位 50 件の中に別のフィッシングと思われるサイトが出現していた．もし，この別サイトが検査対象であったならば，フィッシングを見逃す可能性がある．

マーク C がついているものは，Google/m で正規サイトの検索に失敗し，フィッシングらしき別サイトも検索されなかった．正規サイト検索失敗の代表例として ID17764 について詳しく説明する．このときのキーワードは，携帯サイト誘導ページに共通の特徴のない単語(″登録″ ″メール″ ″利用″ ″換金″ ″送信″)である．これが選ばれたおかげで，ブランド名を含む「モバゲーポイント」という単語が 9 位に落ちてしまい，キーワードに入らなくなってしまった．その結果，重要なキーワードがなくなり，一般的な単語が増えて正規サイトが検索されなかった．ID17765，17766，17980，18490，18508，18676，19721 等も同様であった．

5.6 考察

フィッシングサイト（モバゲーポイント）が正規サイトの模倣でない可能性がある．正規のモバゲータウンにはモバゲーポイントというページがないようである．携帯のサイトは PC サイトに比べてページの構成が簡易なため, 正規サイトを模倣しなくてもフィッシングサイトを簡単に作れることが原因と考えられる．

検査対象のフィッシングサイトと ECHCO システムが検索した別のフィッシングサイトを比べると，コンテンツが同じである．つまり，人を騙すためのモバゲーポイントというページが複数あり，これらはお互い似ているが正規サイトには類似していない．同じ攻撃者が手作りのフィッシングサイトを複数の URL に立てていると考えられる．

5.7 改良案

今回の実験で正規サイトを検索できなかった 8 件について改良案を検討した．

5.7.1 タイトルキーワード手法の導入

ページタイトルをキーワードとすることで，表 3 で示したように，8 件中 6 件は正規サイトを検索できた（ID17764, 17765, 17766, 18490, 18508, 19721）．しかし，検査対象とは異なるフィッシングサイトが 50 位以内に入ってしまうケースが 5 件（19721 以外，全てモバゲーポイント）あったので，フィッシングの検知率が下がると考えられる．よって，タイトルキーワードを単純に導入しただけでは問題は解決しない．

5.7.2 キーワード抽出の高度化

試しに，ID17764 のタイトルキーワードである「モバゲーポイント」をブランド名「モバゲー」にしたら，正規サイトの検出に成功し，別のフィッシングサイトが 50 位以内に出現しなくなった．しかも，Yahoo!API の検索結果にも，正規の PC サイトが出現するようになった．正規の PC サイトは正規の携帯サイトと同一ドメインにあるので，結果的に，Yahoo!API を用いて正規サイトの検索ができ，これまでの ECHCO システムを変更しなくてもいい．

ixen，mixi，gree についても，ブランド名をキーワードとして，Yahoo!API を用いて実験した．結果を以下に示す

- mixi：正規の PC サイトが 7 位で出現
- ixen：正規の携帯サイトが 5 位で出現（PC では開けないので中身は確認できなかった）
- Gree：正規の PC サイトが 10 位で出現

以上から，Google/m の API を利用できるか，または，携帯サイトの少ないテキストからブランド名を正しく取り出せるようにキーワード抽出アルゴリズムを高度化することができれば問題を解決することができる．

6. 新たに提供を受けたＰＣ向けフィッシングサイトへの適用評価

6.1 目的

JPCERT/CC から新たに提供を受けたフィッシングサイトに対して，ドメインキーワード手法を導入した ECHCO システムを適用する．

6.2 実験方法

6.2.1 概要

吉浦研究室が開発したドメインキーワード手法導入の ECHCO システム(ECHCO2 システム)を，(1)JPCERT/CC の保有するフィッシング実例データ 293 件に適用し，その結果を分析する．また，今回の実験では，前年度の実験での 6 通りのモードの中で 1 番よかった②のモードによって実験を行った（表 4）．

表 4 実験モード

タグ除去 / ダイアクリティカルマーク除外	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

6.2.2 評価項目

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

正規サイト導出率：フィッシングサイトを検査し，正規サイトを検索できた率

6.3 実験結果

6.3.1 フィッシングサイト 293 件

293 件のフィッシング実例データを分類したものを図 5 に示す．

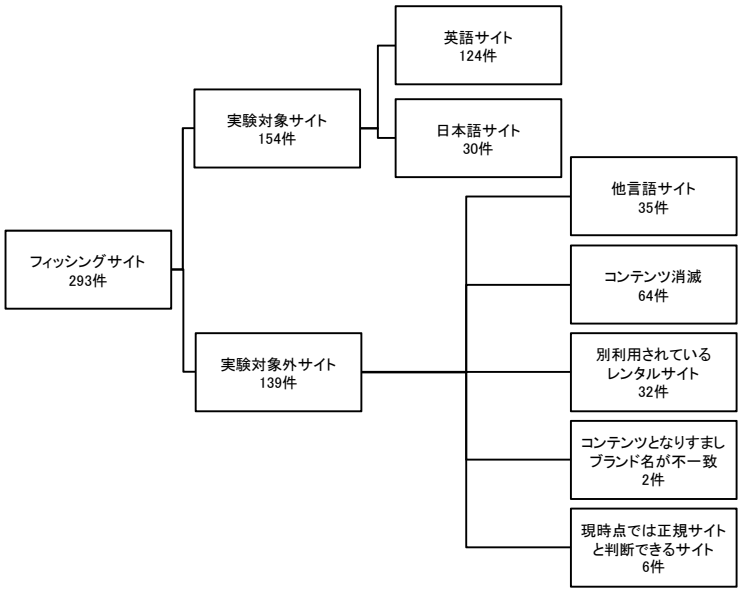

図 5 フィッシング 293 件における実験適用データの分類

他言語サイトについては，対応していないため実験対象外としている．

コンテンツ消滅したページについては，コンテンツベース方式を用いている ECHCO シ

ステムでは対象外である．

別利用されているレンタルサイトとは，実際に稼働しているウェブサイトであるが，ブランドページと違っている．また，これらのサイトが正規サイトかフィッシングサイトかが判断できない．これらのサイトに対しては，今現在実害がないという判断が出来るので，実験対象外とした．

コンテンツとなりすましブランド名が不一致とは，データとして渡されたブランド名が MasterCard であったのに対して，フィッシングサイトは Paypal であった．これは判定をすることができなかったので，実験対象外とした．

現時点では正規サイトと判断できるサイトとは，実際に稼働しているウェブサイトであるが，ブランドページと違っている．そして，これらのウェブサイトはテニスのオンラインショップなどの正規サイトであると判断できた．これらのサイトに対しては，実害がないと判断できるので，実験対象外とした．

6.4 検知結果

6.4.1 概要

検査対象データ 154 件中，154 件全てをフィッシングサイトであると判定できた．以下の図に示す(図 6)．検知結果の詳細については後述する．

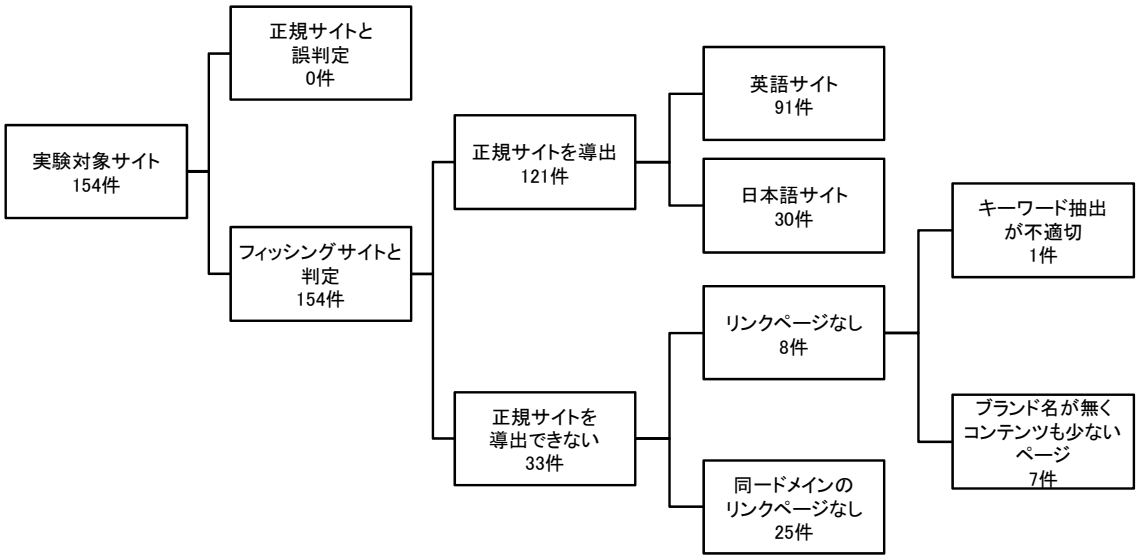

図 6 検知結果の全体図

従来の ECHCO システムにおける正規サイト導出率は 154 件中 85 件であった(55.2%)．そして，ECHCO2 システムは ECHCO システムよりも 36 件多く正規サイトを導出できた．その結果，全体で 121 件の正規サイトを導出できた(正規サイト導出率は 78.6%)．

6.4.2 詳細

詳細な分析結果を表 5 に示す．実験結果は，表 5 の 1 列目（分類）に示すように，A,B,C の 3 パターンに大別される．

A：検査対象ページにリンクがなく，キーワード抽出が不適切なもの．

B：検査対象ページにリンクがなく，ブランド名もなく，コンテンツが少ないもの．

C：検査対象ページに同一ドメインのリンクがないもの．

表 5 のその他の列は下記を示している．

- ID：実験サンプルの番号
- ブランド名：模倣元の企業名
- ECHCO2 のキーワード:ECHCO2 システムが見つけたキーワード(TF-IDF 値)の上位 5 件を記載
- キーワードによる検索結果：ECHCO2 システムが見つけたキーワードを元に検索をした順位の結果
- リンクページの有無：フィッシングサイトにリンクページがあるかどうか
- 理由：検知結果の理由を記載

表5 実験結果のまとめ

分類	ID	ブランド名	ECHCO2のキーワード (従来のECHCOシステムと同様)	キーワードによる 検索結果	リンクページの有無	原因
A	38706	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	38723	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	38730	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	38733	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	38739	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	38850	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	39107	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	39120	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	39121	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	43498	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	43502	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	43519	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	43521	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	43530	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	43569	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	43627	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	46369	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	46406	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	46414	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	46440	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	46454	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	46465	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	46473	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
A	46499	MasterCard	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている
C	42459	Paypal	login account address go password	50位内になし	リンクページなし	ブランド名なし、コンテンツも少ないログインページ
C	42662	Paypal	login account address go password	50位内になし	リンクページなし	ブランド名なし、コンテンツも少ないログインページ
C	43108	Paypal	login account address go password	50位内になし	リンクページなし	ブランド名なし、コンテンツも少ないログインページ
C	43285	Paypal	login account address go password	50位内になし	リンクページなし	ブランド名なし、コンテンツも少ないログインページ
C	43364	Paypal	login account address go password	50位内になし	リンクページなし	ブランド名なし、コンテンツも少ないログインページ
C	43496	Paypal	login account address go password	50位内になし	リンクページなし	ブランド名なし、コンテンツも少ないログインページ
C	45861	Paypal	login account address go password	50位内になし	リンクページなし	ブランド名なし、コンテンツも少ないログインページ
B	43969	TruPoint Bank	trupoint abcd bank inconvenience password	0件	リンクページなし	大文字や小文字の説明を単語と見なした
A	44041	Visa	browser support page	50位内になし	同一リンクページなし	ほとんどがJavaScriptで書かれている

6.5 分析

今回の実験で正規サイトが導出できなかった 33 件について分析する．

6.5.1 検査対象ページにリンクがないもの

リンクがないものは 8 件あった．ドメインキーワード手法は，検査対象ページにリンクが張られているという前提で行うものなので，これに対しては検索できない．

その中で，以下の 2 つに分類できる．

6.5.1.1 キーワード抽出が不適切（A）

表 5 の ID43969 が該当する．具体的には TruPoint Bank というブランドになりすましたフィッシングサイトで，キーワードが trupoint abcd bank inconvenience password であった．この中に”abcd”というキーワードが入っており，このせいで正規サイトが導出できなかった．これは，コンテンツ内で大文字(ABCD)や小文字(abcd)の説明をしていたために，特徴的な語として”abcd”がキーワードに入ってしまった．このような場合でも，適切なキーワードを抽出できるように改良するべきである．

6.5.1.2 検査対象ページにブランド名が無く，コンテンツも少ない(B)

ID42459, 42662, 43108, 43285, 43364, 43496, 45861 の 7 件が該当する．これは人間でも判断できないため対応できない．しかし，正規サイトではこのような状態はないため問題はない．

6.5.2 検査対象ページに同一ドメインのリンクがないもの(C)

同一のドメインのリンクがないものは 25 件あった（表 5 の ID38706 から 46499）．ドメインキーワード手法は，同じドメインのリンクに対してコンテンツ情報を抽出し，単語の出現率を単語の特徴度にかける手法である．その為，同じドメインのリンクがないページに対しては従来の ECHCO システムと変わらない．

また，これらのページでは HTML ソース内にテキストが殆どなく，JavaScript プログラムがユーザ閲覧時にブラウザ上にテキストを表示する形態である．そのため，HTML ソースからキーワードを抽出する現在の ECHCO 方式では，キーワード抽出が困難であるいう問題があった．今回は検査対象がフィッシングサイトであったので，正規サイトが導出できなくても実害は生じないが，検査対象が正規サイトであれば誤検知につながる．そこで，この形態の正規サイトが多いのであれば，ブラウザ表示の中からキーワードを抽出するといった改良が必要となる．

6.6 考察

今回の実験で正規サイトの導出ができなかったもののうち，B については，検査対象が正規サイトの場合に生じない問題なので，対応する必要はない．しかし，A については，大文字小文字の説明における”abcd”のような文字列をキーワードとして抽出することがないように，今後，キーワード抽出アルゴリズムを改良する．また，フィッシング検知の対象となる正規サイトのなかに，プログラムがテキストを生成するような形態が多く含まれるかを調査し，多く含まれる場合には，この形態への対応を行なう．

7. 結論

本研究ではコンテンツベース方式の実用化に向けて正規サイト，携帯向けフィッシングサイト，最新 PC 向けフィッシングサイトを用いた評価実験を行った．

正規サイト 100 件のうち 9 件についてフィッシングと誤検知した．これらの原因を分析した結果，HTML からのコンテンツ抽出手法の改良，N グラム法を用いたキーワード抽出の高度化，検査対象ページだけでなくその周辺ページを考慮するキーワード抽出の高度化（ドメインキーワード手法）によって解決可能であることが分かった．

17 件の携帯向けフィッシングサイトを用いた評価では，17 件全てに対して正しくフィッシング検知ができた．一方，模倣元である正規サイトを検索できたのは 1 件のみであり，16 件については検索できなかった．16 件のうち 7 件については，正規サイトの誤検知にはつながらないが，残りの 9 件についはて正規サイトの誤検知につながる可能性がある．正規サイトを検索できない原因は，現在用いている Yahoo!API では携帯の正規サイトを検索できない点があげられる．しかし，ブランド名を正しく特定できればこれらの正規サイトを検索できることがわかった．そこで，今後の改良として，コンテンツの少ない携帯サイトからブランド名を正しく特定する，キーワード抽出の高度化があげられる．

最新の PC 向けフィッシングサイトへの適用評価を行う前に，正規サイト導出率の向上の最も有効な方法と考えられるドメインキーワード手法を実装し，これまでのコンテンツベース手法に組み込んだ．

最新の PC 向けフィッシングサイトへの適用評価では，154 件のフィッシングサイトを評価し，全て正しくフィッシング検知できた．一方，模倣元の正規サイトを正しく導出できたのは 121 件であった．正規サイトを導出できなかった 33 件のうち 32 件については，正規サイトの誤検知につながらないが，残りの 1 件については対策が必要である．この 1 件では，大文字と小文字の説明の中で，「abcd」という文字列があり，これをキーワードとして選択したことで，正規サイトが検索できなくなっていた．そこで，対策として，キーワード抽出の高精度化が必要である．

今後の課題として以下の改良があげられる．

１．タグ除去時の正規表現の改良
２．N グラム法の導入によるキーワード抽出の向上
３．少ないテキストからキーワード抽出を行う手法の考案
４．誤ったキーワードの抽出を避ける手法の考案
５．プログラムがテキストを生成する形態の正規サイトの調査と対応

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 ITmedia,"サイバー犯罪者の重点投資", http://www.itmedia.co.jp/enterprise/articles/1101/21/news066.html(2011 月 3 月確認)

3 Yue Zhang, Jason Hong, Lorrie Cranor, “CANTINA: A Content-Based Approach to Detecting Phishing Web Sites”, WWW2007, (2007).

4 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).

5 http://www.google.co.jp/m