情報処理学会論文誌　Vol. 50　No. 9　2034–2047 (Sep. 2009)

模倣コンテンツの特性に基づくフィッシング検知方式の誤検知防止

中山心太[†1]　吉浦裕[†2]

コンテンツベースのフィッシング検知方式では，検査対象ページから自然言語処理技術を利用してキーワードを抽出し，そのキーワードを用いてインターネット検索することで，正規サイト候補を得る．そして，検査対象ページと正規サイト候補のドメインと比較することで，検査対象ページの真偽を判定する．しかし，キーワード抽出の精度が不十分なため，検査対象が正規であっても，フィッシングと誤検知することが多い．そこで本論文では，検査対象ページだけでなく周辺ページを含んだ解析をすることで，より広域のキーワードを抽出するドメインキーワード手法と，閲覧サイトの過去のページを参照することで時間的に安定なキーワードを抽出する時間不変キーワード手法を提案し，実装した．正規サイト 172 件を用いた評価実験では，従来手法では誤検知率が 14.0%だったが，提案手法により 7.6%に改善された．フィッシングサイト 172 件を用いた評価実験では，従来手法では 2.9%が誤検知したが，提案手法による誤検知率の増加はなかった．

Preventing False Positives in Phishing Detection Based on Features of Mimic Content

Shinta Nakayama[†1] and Hiroshi Yoshiura[†2]

Content-based phishing detection extracts keywords from the target Web page, uses these keywords to retrieve the corresponding legal site, and detects phishing when the domain of the target page does not match that of the retrieved site. It often judges a legal target page as phishing, however, because the extracted keywords are not adequate. This paper describes two methods of extracting keywords: domain keyword extraction that extracts keywords from not only target page but also other pages linked from the target and time-invariant keyword extraction that extracts keyword from the intersection between the target and its past data. Experimentation using 172 legal pages has shown the decrease of false detection from 14.0% to 7.6% and that using 172 phishing pages has shown that the ratio of overlooking phishing does not have changed.

1. はじめに

子供や高齢者などコンピュータリテラシの低い層のインターネット利用が一般化してきた．これにともない，低リテラシ層をターゲットに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を装ったウェブサイトを制作し，これを用いてユーザからクレジットカード番号や預金口座の暗証番号などを受け取り，金銭を引き出す詐欺である．2006 年度の全米被害額は 28 億ドル，2007 年度は 32 億ドルと年々増加しており[1]，対策は急務である．

既存の対策手法として，正規サイトを列挙したホワイトリストを用いた方式，フィッシングサイトを列挙したブラックリストを用いた方式などがある．しかしこれらの手法はデータベースの頻繁な更新が必要である．

そこで，データベースの更新が不要な，コンテンツベース方式[2),3)]が提案されている．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し，フィッシング詐欺検知を行う方式である．しかしコンテンツベース方式は正規サイトをフィッシングサイトだと誤検知してしまう率が高いという問題がある．

そこで本研究では正規サイトの誤検知を減少させるために，ドメインキーワード手法と，時間不変キーワード手法の 2 つを提案し，実装，評価する．

2. 先行研究

2.1 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方式である[4)]．ホワイトリスト方式では，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．

†1 電気通信大学大学院電気通信学研究科人間コミュニケーション学専攻
The Department of Human Communication, The Graduate School of Electro-Communications, The University of Electro-Communications
†2 電気通信大学電気通信学部人間コミュニケーション学科
The Department of Human Communication, The Faculty of Electro-Communications, The University of Electro-Communications

2.2　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方法である[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，初めて登録される．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．

2.3　ネットワークの性質に基づいた方式

データベースを用いない手法としては，フィッシングサイトのネットワーク的特性を利用したものがある[6]．米国の APWG（Anti Phishing Working Group）[7] の調査によると，フィッシングサイトの平均存続期間は 3.1 日と非常に短い[8]．そのため，ウェブサイトの存続期間に基づいて，ドメインの登録日時，DNS の逆引きが可能かどうか，Google の PageRank なども判定基準として用いることで，フィッシングサイトか否かの判定を行うことができる．しかし，個人サーバや新たにできたウェブサイトは，フィッシングサイトとネットワーク的特性が類似しており，フィッシングサイト扱いしてしまう可能性がある．

2.4　ユーザの認知能力の分析

被験者にウェブサイトを見せてフィッシングサイトかどうか判定させる実験[9] によると，23%の被験者はウェブの内容しか見ておらず，アドレスバーや SSL の錠前のアイコンなどは見ていなかった．また，多くの被験者は SSL の警告メッセージの意味を理解しておらず，最も精巧にできたフィッシングサイトでは 9 割の人を騙すことに成功した．そのため，ユーザの認知能力には限界があるため，技術的な対策が重要であることが分かる．

2.5　視覚的類似性に基づいた方式

フィッシングサイトと正規サイトが視覚的に類似しているという仮説のもとに，視覚的類似性を判定することでフィッシング検知をする[10]．しかし HTML のタグ情報を解析して，デザイン情報の類似性を判断しているので，タグ情報を書き換えることで，容易に検知を逃れることができる．また，疑わしいサイトと，正規サイトとの両方が与えられていることを前提とした判定方法であるため，正規サイトの特徴情報を保存したデータベースが必要になる．

2.6　コンテンツベース方式

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．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いる．フィッシングサイトの多くは正規サイトをコピー，もしくは模倣したものである．そこで，コンテンツの類似性を利用したコンテンツベース手法が Zhang ら[2]，中山ら[3] によって提案されている．

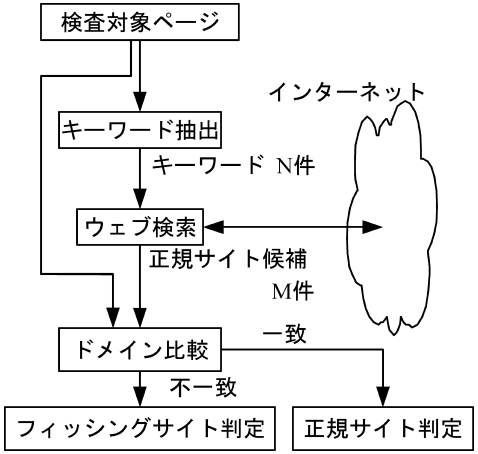

図 1　コンテンツベース方式の処理の流れ
Fig. 1　Flowchart of content-based method.

フィッシングサイトを熟知したユーザは，フィッシングサイトの疑いのあるウェブページを見た際，そのウェブページの特徴を現す語句（企業名，製品名など）をキーワードにして，ウェブ検索を行うことがある．フィッシングサイトの存続期間は 3.1 日と短く[8]，また他のウェブサイトからリンクされることが稀であるため，検索エンジンからの評価が低い．そのため，適切なキーワードを選べば正規サイトのみが検索結果に現れ，フィッシングサイトは現れないというようになる．そこで，疑わしいウェブページの URL と，検索結果に現れた URL を比較することで，そのウェブページがフィッシングサイトかどうかを判断することができる．

この手法を計算機上で再現したのが，コンテンツベース方式である．この方式は，検索エンジンをある種のホワイトリストとして利用するため，ブラックリストやホワイトリストといったデータベースを持つ必要がないという特徴がある．

コンテンツベース方式の処理の流れは**図 1** のようになっている．検査対象ページを入力すると，自然言語処理技術によってウェブページを最もよく表すキーワード（会社名や製品名）を抽出する．次にこのキーワードを用いて検索を行う．これにより，そのキーワードに

最もよく適合するウェブページが，正規サイト候補として検索される．

もし検査対象が正規サイトであった場合，検索された正規サイト候補の中に検査対象サイトが含まれるはずである．そのため，検査対象サイトと正規サイト候補のドメインを比較することで，ドメインが同一であれば正規サイト，一致していなければフィッシングサイトであると判断することができる．

なお，キーワード抽出には TF-IDF 法[11] が用いられている．TF-IDF 法は文章中の単語の特徴度を計算する手法で，式 (1) で表される．ドキュメント d 中の単語 t の特徴度 w を，ドキュメント d 中の t の出現回数 $tf(t,d)$ と，サンプルドキュメント集合中に t がどれほど現れているか，という単語の稀少性 $idf(t)$ の積によって定義する．また，D はサンプルドキュメントの総数，$df(t)$ はサンプルドキュメントのうちに t を含むドキュメントの数を表す．

$$w(t,d) = tf(t,d) \cdot idf(t)$$
$$idf(t) = \log\left(\frac{D}{df(t)}\right) \quad (1)$$

3. コンテンツベース方式の課題と改善手法

コンテンツベース方式には正規サイトをフィッシングサイトと誤って判断してしまう率が高いという問題がある．そこで，いくつかの誤検知防止手法が提案されている．

3.1 経験則による改善

Zhang らの CANTINA 方式[2] では，コンテンツベース方式に加えて，以下の経験則を併用してウェブサイトの疑わしさのスコアを算出し検知精度を向上させている．

① ドメイン登録から 1 年以下か．
② 既知の画像を使っているか．
③ URL 中に「@」や「-」を含んでいるか．
④ ページ中のリンクの URL に「@」や「-」を含んでいるか．
⑤ ドメインネームを使わずに IP を使っているか．
⑥ URL 中にドットが 5 個以上含まれているか．
⑦ ページ中にテキスト入力フォームが存在するか．

CANTINA 方式では，これらの経験則を導入することで，正規サイトの誤検知を 6%から 1%に減少させた．しかしフィッシングサイトの検知率は 97%から 89%に低下している．また，経験則の判断基準をフィッシングサイト製作者が知っていれば，それらを回避したフィッシングサイトを作成することは容易であり，89%よりも大幅に検知率が下がることが予想される．

表 1　コンテンツ一致によるフィッシングサイト判定方法
Table 1　Detection of phishing site by content coincidence.

	コンテンツが一致	コンテンツが不一致
ドメインが一致	①正規サイト	②正規サイト
ドメインが不一致	③フィッシングサイト	④検索失敗

3.2 Web of Trust による改善

ASP でウェブ業務の一部を外部委託している場合がある．たとえば国内では地方銀行のオンラインバンク業務を代行している anser.or.jp などが有名である．このときある企業のウェブサイトは，企業自身のウェブサイトと，外部委託先という 2 つのドメインで運用される．そのため外部委託先をコンテンツベース方式で検査すると，正規サイト候補の中に企業自身のウェブサイトは含まれるが，検査対象のドメインは含まれないというケースが発生する．

たとえば A 銀行がオンラインバンク業務を別ドメインの他社に業務委託しているとする．A 銀行のオンラインバンクページ（業務委託先ドメイン）をコンテンツベース方式で検査すると，得られるキーワードは A 銀行の特徴を現したものであり，このキーワードでウェブ検索を行うと，A 銀行のウェブサイトが正規サイト候補として得られてしまう．そのため，ドメインが異なるため，フィッシングサイトと誤検知してしまう．

そこで西垣らは Web of Trust による改善を行った[12]．Web of Trust とは「多くの人が信頼しているものは信頼できる」というモデルである．このとき正規サイト候補は正しいと仮定し，正規サイト候補からリンクをたどり，検査対象のサイトにたどり着ければ，正規サイトであると判定する．前述の例では，正規サイト候補である A 銀行のウェブサイトからリンクをたどり，オンラインバンク機能の業務委託先のドメインにたどり着ければ，正規サイトであると判断する．

3.3 コンテンツ一致方式による改善

コンテンツ一致方式は中山らが提案[13] している．**表 1** の評価基準を用いて，正規サイト候補のドメインだけでなく，コンテンツを比較することで，正規サイトの誤検知を検出する手法である．従来はドメインネームの比較を行うのみであるため，ドメインが一致していない場合，無条件でフィッシングサイトと判断していた．しかし，ドメインが不一致した際に

コンテンツを比較することで，フィッシングサイトなのか，何らかの原因で検索が失敗したのかを知ることができる．それぞれのケースは次のようになる．

① 自分自身を検索できたことから正規サイトであると判断する．
② 正規サイトの別ページが検索されてしまっていると考えられる．そのため正規サイトであると判断する．
③ 正規サイトとフィッシングサイトとで利用される言葉は，ユーザを騙すためにほぼ同一であり，コンテンツは一致している．そのためフィッシングサイトであると判断する．
④ 得られた正規サイト候補がまったくの別ものであると考えられる．そのため，検索キーワードの選定に失敗し，その結果正規サイトの検索に失敗したと判断できる．

4. 正規サイトの誤検知の発生原因の分析

正規サイトの誤検知の発生原因は，著者らが過去の研究会発表において分析している[19]．

4.1 検索エンジンに載らないページ

robots.txt や meta タグをウェブページに付加すると，そのウェブページは検索エンジンの検索対象から除外される．これを利用すると，オンラインバンキングなどのページをユーザに直接検索させず，まずウェブサイトのトップページを検索させ，トップページの情報を閲覧させたうえでオンラインバンキングを利用してもらうことができる．しかし，検索エンジンに載らないページが検査対象ページであったとき，検査対象ページ中のキーワードを利用しても，自分自身を検索することができないのでフィッシングサイトと誤判定される．

4.2 複数ドメインで運営されているサイト

子会社を持っている企業や，様々なサービスを運用しており，サービスごとに独自のドメインを持っているような企業の場合，運営母体が同一であるため利用されている単語は非常に似通っている．そのため，検査対象ページを自然言語処理し，得られた単語で検索しても，同一運営母体の別ドメインの正規サイトが検索されてしまう．これにより，自分自身を検索できずフィッシングサイトと判定されてしまう．

4.3 自然言語処理の失敗

現行の検索エンジンの多くは，全文検索ではなく，インデックス検索である．そのため，文中に存在する文字列であっても，インデックスに載っていない単語は検索できない．実際にあった例では，HTML のパースに失敗して，JavaScirpt のコードの一部がキーワードとして選ばれてしまったり，2 語が 1 語に結合されてしまい，新しい単語が生まれてしまったというケースがあった．

また，このような自然言語処理の失敗によって生まれた単語は稀少であり，他のページで利用されていることが少ない．そのため，TF-IDF 法では高い評価値を得ることになり，キーワードとして非常に選ばれやすい．

そのため，自然言語処理の失敗によって生まれた単語を検索キーワードとして選んでしまい，正規サイト候補の検索に失敗し，フィッシングサイトと誤判定をする．

4.4 検索エンジンのキャッシュとの相違

検索エンジンはウェブサイトを定期的に巡回し，ウェブサイトの情報を解析し，検索エンジンのインデックスに収めていく．そのため，最新のウェブサイトと，検索エンジンが解析したウェブサイトとが異なる場合がある．その際，最新のウェブサイトには存在するが，検索エンジンのキャッシュに存在しない言葉をキーワードとして選んでしまった場合，検索に失敗し，フィッシングサイトであると判断されてしまう．

5. 改善手法の提案

従来のコンテンツベース方式[2),3)] は検査対象ページから，自然言語処理でページの特徴を表すキーワードを得て，そのキーワードで正規サイトのページを検索する．検索結果で得られるページは，検査対象ページ自身ではなく，正規サイト内の別ページであることが多いため，検査対象ページと正規サイト候補とは，ドメインレベルの比較をする．これは，検査対象ページのキーワードを用いて，正規サイト内のどれかのページを検索しようとしていることになる．すなわち，ウェブページではなく，ウェブサイトを検索していると考えられる．

ウェブサイトを検索するためには，検査対象ページだけでなく，それを含むウェブサイトの特徴を表すキーワードを用いるべきである．ウェブサイトのキーワードを用いることにより，正規サイトが正しく検索される可能性が高まり，4 章で分析した 4 種類の誤検知の問題を軽減させることができると考える．

ウェブサイトのキーワードを得るための手法として，ドメインキーワード手法と，時間不変キーワード手法という 2 つのキーワード選定方法を提案する．

5.1 ドメインキーワード手法

ドメインキーワード手法は，検査対象ページから同一ドメイン内のリンクをたどり，自然言語処理によるキーワード抽出の対象をウェブサイト全体に広げる手法である（**図 2**）．そして，リンク先で利用されている言葉の頻度を調べることにより，ウェブページではなく，ウェブサイト全体に特徴的な単語をキーワードとして得る．これにより，前述の 4.1，4.2，4.3 節の問題を解決する．

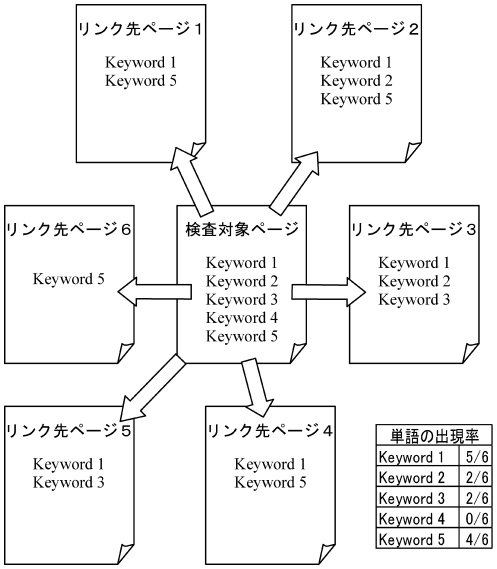

図 2　ドメインキーワード手法
Fig. 2　Domain keyword method.

5.1.1　検索エンジンに載らないページへの対策

検査対象ページの特徴のみを表し，ウェブサイトの特徴を表さない単語をキーワードとして選んだ場合，ドメイン内の他のページは検索されない．さらに検査対象ページは検索対象から除外されているので，検索されない．そのため，同一ドメイン内のページがまったく検索されず，フィッシングサイトと誤検知される．たとえば図 2 中の Keyword 4 のような単語をキーワードとして選んでしまうと，検査対象ページにおいては特徴的な言葉であっても，ウェブサイト全体ではまったく出現せずウェブサイトにとって特徴的な単語とはいえない．そのため，Keyword 4 では同ドメインのウェブページを検索することができず，フィッシングサイトであると判定してしまう．しかしリンク先のページをたどり，単語の出現率を調べることで，Keyword 4 のような単語は排除され，ウェブページ全体に共通して見られる Keyword 1 のような単語が選ばれるようになる．その結果，同ドメインのページを検索可能になる．

5.1.2　複数ドメインで運営されているサイトへの対策

同一運営母体のウェブサイトにおいて，利用されている単語はよく似通っている．たとえば会社 A が子会社 B を持っているとする．また検査対象ページが子会社 B のウェブサイトの 1 つのページであったとする．このとき B は A の子会社であるため，B のウェブページには A に特徴的な単語と B に特徴的な単語との両方が存在する．もし A に特徴的な単語を検索キーワードとして選んでしまうと，一般に B よりも A のほうが検索エンジンの評価が高いため，A のウェブサイトが検索され，フィッシングサイトであると判定されてしまう．

しかし B のウェブサイトには，B に特徴的な単語が A に特徴的な単語よりも多く含まれているような場合も多いと考えられる．そのようなウェブサイトに対しては，ドメインキーワード手法を利用し，リンク先のページで利用されている単語を調べることで，B に特徴的な単語が検索キーワードとなる可能性を高めることができる．その結果，B ドメインのページが正規サイト候補として検索される可能性が高まり，正規サイトと判定されやすくなる．

5.1.3　自然言語処理の失敗への対策

自然言語処理が失敗して，たとえば 2 つの言葉が 1 つに結合したりすると，図 2 中の検査対象ページには本来含まれていない単語が生まれてしまう．この本来含まれていない単語を検索キーワードに使って検索を行っても，検査対象ページは検索されない．また，そのような単語は他のウェブサイトに出現することも稀であるため，TF-IDF 法による特徴度は高くなってしまい，キーワードに選ばれやすい．そこで，そのような自然言語処理の失敗によって生まれた単語を取り除く必要がある．

自然言語処理が失敗する確率は低いため，リンク先のページで同様の自然言語処理の失敗が起こることは稀である，そのため，リンク先のページで使われている単語の出現頻度を参照することで，そのような単語を除去することができる．

5.1.4　ドメインキーワード手法の限界

ドメインキーワード手法は，検査対象ページ中に同一ドメイン内への多くのリンクが存在することが前提である．そのため，末端のウェブページなどではリンク先ページが存在しないため，ドメインキーワード手法を利用することはできない．また，リンク先ページが少ない場合には，適切なキーワードが選択されない可能性がある．たとえばリンク先のページが 1 件しかなく，共通して利用されている単語がない場合，ドメイン内で共通して利用されている単語を選ぶことができない．

このようにドメインキーワード手法はつねに有効とは限らない．しかし，正規サイトを正

しく検索できる可能性を高め，誤検知率を低減させることができると考えられる．ドメインキーワード手法の有効性については，7 章の評価実験により確認する．

5.2　時間不変キーワード手法

時間不変キーワード手法は，4.4 節の問題を解決する．検索エンジンは 1 日に数回ウェブページを巡回し，ウェブページのキャッシュを保存し，検索のためのキーワードリストを作成する．また，ウェブページは新たなニュースが掲載されたり，コンテンツが掲載，削除されたりすることによって，時々刻々と変化する．そのため，検査対象ページには存在しているが，検索エンジンのキャッシュには存在しない単語というものが存在する．たとえば**図 3** 中の Keyword 5 のような単語をキーワードとして選定すると，自身を検索することに失敗してしまう．

Keyword 5 のような変化した部分のコンテンツは，その時点のウェブページの特徴を表す情報ではあるが，ウェブサイトの特徴を表す情報ではない．そこで，検査対象ページの過去のページをいくつか取得し，検査対象ページに含まれる単語集合と，その過去ページ群に含まれる単語集合の積集合をとる．このようにすることで，検索可能でなおかつウェブサイトの特徴を表すキーワードを得ることができる．図 3 の例では Keyword 1 のみが過去ページのすべてで現れているため，この単語が検索キーワードとして選ばれる．検索エンジンのキャッシュには Keyword 1 が含まれるため，自身を検索することができ，正規サイトであると判断することができる．

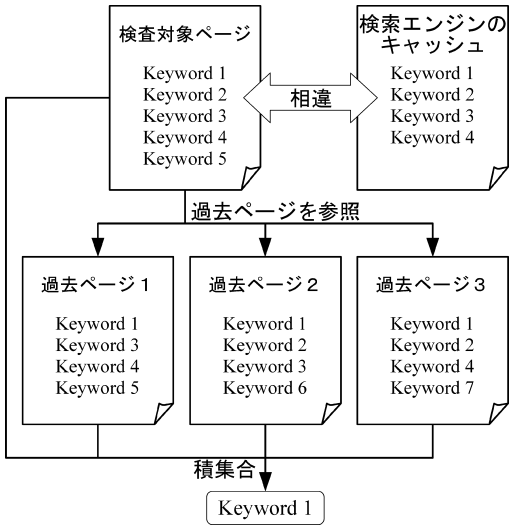

図 3　時間不変キーワード手法
Fig. 3　Time-invariant keyword method.

6. 実　　装

6.1　ドメインキーワードの実装

ドメインキーワードの実装は，いくつか方法が考えられるが，今回は式 (2) のように実装した．

$$w(t,d) = tfidf(t,d) \cdot \frac{df_L(t)}{L} \tag{2}$$

$tfidf(t,d)$ は式 (1) による TF-IDF 法による特徴度，L は検査対象ページからの同一ドメイン内のリンク先ページの総数，$df_L(t)$ はリンク先のページのうち t を含むページの数を表す．TF-IDF 法で特徴度を計算したものに，リンク先での言葉の出現率を掛け合わせている．これにより，他のページにほとんど現れない，そのページ固有のキーワードが排除される．

ドメインキーワード手法を実装し，予備実験を行った．結果を**表 2** に示す．予備実験では代表的なウェブサイトのトップページをサンプルとし，TF-IDF 法とドメインキーワード手法によって特徴語抽出を行い，各々得られた特徴語の上位 5 件を利用して Google による検索を行った．

(1) オークションサイト eBay.com

eBay.com は動的コンテンツが多く，JavaScript が多用されており，本システムによる

表 2　ドメインキーワード手法の予備実験結果
Table 2　Preliminary experimentation of domain keyword method.

ウェブサイト		eBay.com		yahoo.com		CNN.com	
手法		TF-IDF法	ドメインキーワード	TF-IDF法	ドメインキーワード	TF-IDF法	ドメインキーワード
上位10個の特徴語	1	ebay	ebay	yahoo	yahoo	cnn	cnn
	2	microplace	microplace	news	search	biden	news
	3	wemyss	rent.com	search	news	ireport.com	cnn.com
	4	rent.com	collectibles	accountant	page	news	biden
	5	figur	certificates	irs	home	cnn.com	time.com
	6	collectibles	half.com	saber-toothed	site	crash	crash
	7	old/mot	stubhub	domain	sports	kyrgyzstan	kyrgyzstan
	8	tabby	shopping.com	pharyngitis	web	plane	tv
	9	certificates	shop	games	mail	blitzer	plane
	10	half.com	help	sports	sign	time.com	video
上位5個によるGoogle検索順位		**検索不能**	**1位**	**4位**	**1位**	**1位**	**1位**

HTML のパースに失敗した．その結果，「wemyss」や「old/mot」といった，ウェブページで利用されていない単語がキーワードとして抽出された．TF-IDF 法の上位 5 個の単語でウェブ検索を行ったところ，検索結果が 0 件であり，eBay 自身を検索することはできなかった．しかし，ドメインキーワード手法を用いることによって，他のページでは自然言語処理の失敗が起こらず，同様の単語が存在しないことから，そのような単語が排除された．これにより上位 5 件をキーワードとし，eBay.com を検索することができるようになった．

(2) ポータルサイト Yahoo.com

特徴語の上位 5 個をキーワードに利用し検索を行った．Yahoo.com の正規サイトの検索順位は，TF-IDF 法では 4 位，ドメインキーワード手法では 1 位であった．このことから，ドメインキーワード手法がよりウェブサイトの特徴をとらえているということがいえる．

(3) ニュースサイト CNN.com

TF-IDF 法によって得られたキーワードから，ドメインキーワード手法を適用することによって「ireport.com」が消えた．しかし，それ以外は大きな変化は見られなかった．トップページのニュースのヘッドラインと，リンク先のニュース記事で使われる単語がほぼ同一であるためであると考えられる．

6.2　時間不変キーワード手法の実装

図 3 に従って時間不変キーワード手法の実装を行った．あるウェブページの過去のページを取得するには Internet Archive [14] を利用した．Internet Archive とはウェブページの過去の状態を記録，公開しているウェブサービスである．ここでは Internet Archive によって得られたあるウェブページの過去のページを『過去ページ』と定義する．Internet Archive は，アクセス頻度の高いウェブサイトでは 1 日に数回，頻度の少ないウェブサイトでは数日に 1 回，ウェブページを巡回しコンテンツを保存している．また過去ページを公開するのは，ページが収集されてから約半年以降である．

たとえば，あるウェブページ（http://www.example.com）の過去ページを参照したい場合，http://web.archive.org/web/∗/http://www.example .com という URL にアクセスする．これによって，Internet Archive が保存している当該ウェブページの過去ページのリストを HTML 形式で得ることができる．これを利用することで，検査対象ページの任意の時点の過去ページを得ることができる．

検査対象ページで利用されている単語の集合を S_0 とする．最新の過去ページ n 件で利用されている単語の集合を，新しいものから順に S_1，S_2，$S_3, \ldots, S_n$ とする．S_1 は最も新しい過去ページの単語集合，S_2 はその前の過去ページの単語集合である．そして $A_0 = S_0$ とし，$A_i = A_{i-1} \cap S_n$ とする（$i = 1$〜n）．このとき A_i は S_0 から S_i までの単語集合の積集合を意味する．このように A_i を算出することで，長期間にわたって検査対象ページに掲載され続けている単語を残すことができる．

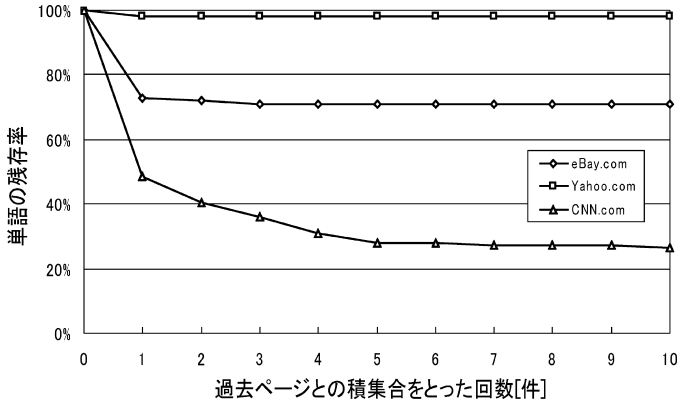

図 4　時間不変キーワードの予備実験結果．件数ベース

Fig. 4　Preliminary experimentation of time-invariant keyword method. In terms of the number of previous pages.

なお，本論文での実装では，HTML を取得するだけで，JavaScript によって動的生成，動的読み込みされる単語は単語集合の中に含めていない．

予備実験としてオークションサイト eBay.com，ポータルサイト Yahoo.com，ニュースサイト CNN.com を取り上げ，各々A_0 から A_{10} を求めた．$|A_i|/|A_0|$ を算出し，A_0 からの単語の残存率とした．ただし $|A_i|$ は A_i に含まれる単語の数である．

図 4 はこの残存率を件数ベースで図示したもの，**図 5** は時間軸ベースで図示したものである．また，実験を行ったのは 2008 年 11 月であるため，それぞれの S_1 は約半年前に保存されたものであるといえる．

(1) オークションサイト eBay.com

過去ページの範囲は 2008 年 3 月 27 日から 2008 年 1 月 30 日の約 2 カ月間にわたっている．

A_1 の時点の単語の残存率は約 73%であり，27%の単語が消えた．A_3 の時点での残存率は約 72%であり．A_4 以降の変化はなかった．

(2) ポータルサイト Yahoo.com

過去ページの範囲は 2008 年 3 月 25 日から 2008 年 2 月 1 日の約 2 カ月間にわたっている．

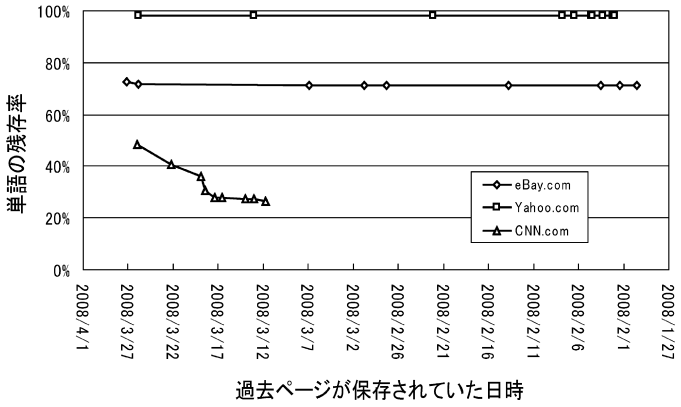

図 5　時間不変キーワードの予備実験結果．時間軸ベース
Fig. 5　Preliminary experimentation of time-invariant keyword method. In terms of time.

A_1 の時点の単語の残存率は約 98%であり，2%の単語が消えた．しかし，A_2 以降はは単語の残存率に変化はなかった．

(3) ニュースサイト CNN.com

過去ページの範囲は 2008 年 3 月 25 日から 2008 年 3 月 11 日の約半月間にわたっている．

A_1 の時点の単語の残存率は約 48%であり，それ以後も減り続け，A_{10} では最終的に 27%の単語が残った．

以上から次のようなことが分かった．

eBay.com では，約 72%が時間の影響を受けにくい単語であり，残りの約 28%は時間の影響を受けやすい単語であることが分かった．InternetArchive が保存していた半年前の S_1 の過去ページと，現在の S_0 のページを比較すると，HTML レベルで大きく異なっていることが分かった．そのため，半年以内にウェブページをリニューアルしていたと考えられる．

Yahoo.com はコンテンツの一部に最新のニュースを表示しているが，ニュースの部分は JavaScript を利用して動的に読み込まれている．本実装では JavaScript によって動的に読み込まれた単語は単語集合に含めていない．そのため，ニュース以外の時間的に安定な部分のみを単語集合としているので，何度積をとっても残存率はほとんど変化しない．

CNN.com では半年前の A_1 以降も単語の残存率は減っていくため，A_1 だけでは，時間の影響を受けやすい単語を除くことができない．これは CNN がニュースサイトで毎日更新されているためであると考えられる．

以上から，時間不変キーワード手法を用いて A_1 を求めることで，約半年間にわたってウェブページに不変な単語を求めることができることが分かった．これにより，たとえばリニューアルなどによって利用される単語が変化した場合，リニューアル前とリニューアル後で共通する単語を選ぶことができる．しかし CNN のような日々変わっていくコンテンツでは，現在のページに含まれる単語がたまたま過去ページにも含まれているというケースが発生する．そのため，直近の A_1 だけでなくより過去のページも調べることで，たまたま含まれてしまうという問題を解決することができ，長期間にわたってウェブページで不変な単語のみを残すことができる．また，Yahoo.com のように動的読み込みをするページや，ほとんど更新しないページでは，効果がないことが分かった．

図 4，図 5 が示すように A_5 で残存率がほぼ収束するため，速度と精度の兼ね合いから提案システムでは A_5 を利用することにした．

6.3　処理の流れ

6.3.1　従来手法の実装

図 1 に従い既存研究[2] 相当の実装を行った．キーワード抽出のための品詞解析エンジンに TreeTagger[15] を利用した．また TF-IDF 法による特徴語抽出は Yahoo! 検索 Web サービスを利用して実装した[16]．IDF の計算にはサンプルドキュメントの総数と，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 サービスが保持しているすべてのウェブページと考え，サンプルドキュメント総数を Yahoo! 検索 Web 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ときのヒット件数とした．

既存研究[2] から，ウェブ検索のためのキーワードは 5 件利用し，またキーワード中にウェブページのドメイン名を含めることとした．また，正規サイト候補は検索結果の上位 10 件を使うこととした．

6.3.2　全体処理の流れ

提案手法を組み込んだ全体の処理の流れを**図 6** に示す．従来手法では，検査対象ページと正規サイト候補とのドメインが一致していない場合，フィッシングサイトと判定する．提案手法を組み込んだ場合，従来手法で検査対象ページと正規サイト候補のドメインが一致していない場合，ドメインキーワード手法で検索キーワードの再選定を行い，上位 5 件のキーワードで改めて検索を行う．それでも得られた正規サイト候補とドメインが一致していない場合，時間不変キーワード手法によりキーワードの再選定を行い，上位 5 件のキーワード

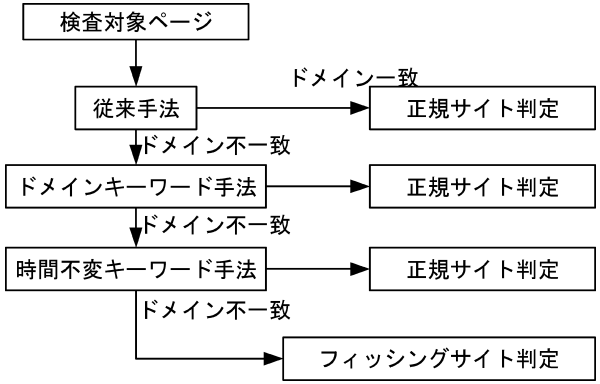

図 6　提案手法を組み込んだ全体の処理の流れ
Fig. 6　Flowchart of total system.

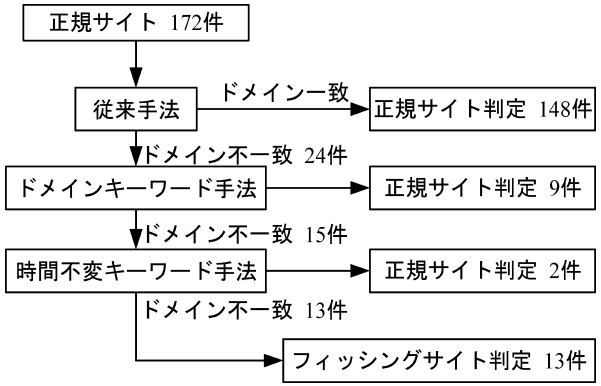

図 7　正規サイトの実験結果
Fig. 7　Experimentation result for legal sites.

で改めて検索を行う．それでもドメインが一致しなかった場合はフィッシングサイトと判断する．

7. 評　　価

7.1 正規サイトの評価

(1) 全体評価

実験のためのウェブサイトは，Yahoo.com の Bank ディレクトリ[17] を人気順（Popularity）でソートしたものから，上位 200 件を利用した．しかし 200 件中，13 件はウェブページを開くことができなかったため除外した．また，品詞解析エンジンは英語を対象にしているため，英語以外の言語を用いた 15 件を除外した．そのため，実質的な母集団は 172 件となる．実験結果を**図 7** に示す．従来手法の誤検知率は 14.0% (24/172) であり，提案手法の誤検知率は 7.6% (13/172) である．

(2) 個別評価

同じデータセットを用いて，検知手法の個別評価を行った．実験結果を**表 3** に示す．従来手法単独の誤検知率は 14.0%である．一方提案手法であるドメインキーワード手法単独では 42.4%，時間不変キーワード手法単独では 35.5%であり，従来手法より劣る．

ドメインキーワード手法の誤検知率が従来手法より高いのは，検査対象ページに特徴的な単語を除いてしまうからであると考えられる．

表 3　検知手法の組み合わせ方による検知結果
Table 3　Experimentation results for combinations of detection methods.

評価手法	検知件数 [件]（母数 172）	検知率 [%]	誤検知件数 [件]	誤検知率 [%]
従来手法	148	86.0	24	14.0
従来手法＋ドメインキーワード手法	157	91.3	15	8.7
従来手法＋時間不変キーワード手法	151	87.8	21	12.2
従来手法＋ドメインキーワード手法＋時間不変キーワード手法	159	92.4	13	7.6
ドメインキーワード手法	99	57.6	73	42.4
時間不変キーワード手法	111	64.5	61	35.5

たとえば Bank & Trust Company の例では，検査対象ページがログインページであった．従来手法では「banking」「trust」「bank」「id」「member」というキーワードを得た．そしてこれらのキーワードを用いることで，検査対象ページを正しく検索することができ，正規サイトと判断することができた．しかし検査対象ページのリンク先のページは会社説明やサービス紹介であり，「id」というログインページ特有の単語は存在しない．そのため，ドメインキーワード手法を導入すると「id」というキーワードが除かれ，代わりに「company」というキーワードが得られた．その結果キーワードがすべて一般的な名詞となってしまい，

これらのキーワードを用いても Bank & Trust Company のどのページも検索することができず，フィッシングサイトであると誤検知した．

時間不変キーワードの誤検知率も従来手法より高い．現時点でこの原因は不明であり，原因の分析は今後の課題である．

ドメインキーワード手法，時間不変キーワード手法は，単独では検知精度は低い．しかし，提案方式では，従来手法でフィッシングとなる場合についてこれらの手法を適用することで，全体として正規サイトをフィッシングと判定する誤検知率を減らすことができている．一方，7.2 節で後述するように，これらの提案手法を追加しても，フィッシングサイトを正規サイトと誤検知する率は本評価実験の範囲では増加しなかった．

7.1.1　ドメインキーワード手法による改善例

(1) Simens Financial Service

Simens Financial Service のページには「7/7/2008 Billions of inefficiently ……」という文章が含まれていた．TreeTagger を用いてこの文章を解析した結果，単語分割に失敗し，「7/7/2008Billions」を 1 つの単語として認識した．そして従来手法では，この単語を検索キーワードに選んでしまい，正規サイトの検索に失敗した．しかし，リンク先のページには同様の文章が存在しないため，同じ単語が誤認識されることはない．そこでドメインキーワード手法を用いることによって，この誤認識の単語を検索キーワードから除くことができ，代わりにサービスの略称である「sfs」をキーワードにすることができた．新しいキーワードを用いて検索することで，提案システムは Simens Financial Service のページを正しく検索することができ，正規サイトであると判断することができた．

(2) Monarch Community Bank

従来手法で得られたキーワードは「covererage」「fdic」「monarch」「business」「atm」であった．このキーワードを用いて検索した結果，正規サイト候補として Monarch Bank という別の銀行のページが得られた．しかし，ドメインキーワード手法により「monarch」「business」「bank」「community」「account」というキーワードを得ることができた．これは会社名である Monarch Community Bank が多くのリンク先ページで現れたことにより，community という単語の出現率が高くなったためである．このキーワードにより，Monarch Community Bank を検索することができ，正規サイトであると判断することができた．

7.1.2　時間不変キーワード手法による改善例

(1) Bienvendo a BLADEX

検査対象ページの HTML 内には次のような form 文があった．

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

この form 文はタグ中の文字列に『<』『>』が使われており，HTML の規約に違反している．そのため，「isEmail(email_user」というタグの一部が HTML パーサによりコンテンツの一部と判断され，さらにキーワードとして選択された．その結果，正規サイトの検索に失敗し，フィッシングサイトであると誤検知した．

当該ページでこのような規約違反の HTML が利用されるようになったのは半年以内であり，過去ページには規約違反の HTML はなかった．そのため，時間不変キーワード手法により，誤ったキーワードを取り除くことができ，正規サイトであると判断することができた．

(2) First State Bank of Wyoming

検査対象ページでは，当日の天気と気温が表示されている．従来手法は「29f」「fsbw」「fdic」「wyoming」「ira」という検索キーワードを選んだ．そのうち「29f」は当日の気温を表す単語であり，これは日によって異なる．この「29f」を検索キーワードにしたため，正規サイトの検索に失敗した．

時間不変キーワード手法によって，天気と気温の情報を除くことができ，「fsbw」「fdic」「wyoming」「ira」「home」というキーワードを得た．このキーワードで検索を行った結果，First State Bank of Wyoming の検索に成功し，正規サイトであると判断することができた．

7.1.3　改善できなかった例

(1) Rabobank

検査対象ページは JavaScript による自動転送のページであった．実験に用いたシステムは HTML の取得のみを行い，JavaScript の実行は行わないので，転送先のページを解析することができず，キーワード抽出に失敗した．

検査対象ページが JavaScript による自動転送ページや，XMLhttpRequest などを用いた動的生成ページであった場合，JavaScript を実行しなければページの特徴となる単語を取得することはできない．この問題は本システムに JavaScript のエンジンを組み込んだり，ブラウザに本システムを組み込んだりすることで解決できる可能性がある．

(2) JUNIPER

検査対象ページでは，サービス名を表す「JUNIPER」が画像として表示されていた．本

システムでは，テキストのみを評価対象として扱うため，「JUNIPER」という特徴的な単語を得ることはできなかった．

しかしリンク先のページでは「JUNIPER」という単語が利用されているため，ドメインキーワード手法を改良することにより，この単語を選定できる可能性がある．

(3) RURAL AMERICAN BANK

検査対象ページは，「www.ruralamericanbank.com」というドメインであった．しかしシステムは「www.fnbpinecity.com」というドメインのページを正規サイト候補として検索した．そしてドメインが一致しないためフィッシングサイトと判断した．

アドレス解決をすると両者は同一の IP アドレスであることが分かった．以上から「www.fnbpinecity.com」と「www.ruralamericanbank.com」のドメインは異なるが同一の IP アドレスであることが分かった．しかも両者のコンテンツはまったく同一であった．

以上から検査対象サイトは複数のドメインで運営されている正規サイトの 1 つのドメインであること，また片方のドメインが検索エンジンで優先的に検索されることが分かった．これにより誤検知が発生していることが分かった．

この問題は検査対象サイトと正規サイト候補のドメインを比較するだけでなく，アドレス解決を行い IP アドレスが一致しているかどうかを調べることで解決可能である．しかし 2 つの正規サイトが同一のコンテンツで，ドメインも IP アドレスも異なっているケースが実在したため，IP アドレスの比較だけでは誤検知を完全に防ぐことはできない．このようなケースに関しては従来手法との併用を検討する必要がある．

7.2　フィッシングサイトの評価

フィッシングサイトを収集，公開している PhishTank [18] から，最新のフィッシングサイト 200 件を利用した．200 件中 28 件はすでにウェブサイトが閉鎖されており，アクセスすることができなかった．そのため実質的な母集団は 172 件となる．実験結果を**図 8** に示す．

従来手法はフィッシングサイト 172 件中 5 件（2.9%）を正規サイトであると誤検知された．提案手法は従来手法と同一の 5 件のページを誤検知した．以上のように本評価実験の範囲内では誤検知率の変化はなかった．

8.　既存の誤検知防止方式との比較

既存研究[2] のうち経験則を用いない部分を基本的コンテンツベース方式と呼ぶことにする．3 章で述べた従来の改善方式および，本論文の提案方式は，いずれも基本的コンテンツベース方式の誤検知率の低減を目的としている．本章では，従来の改善方式と本論文の提案方式を比較する．

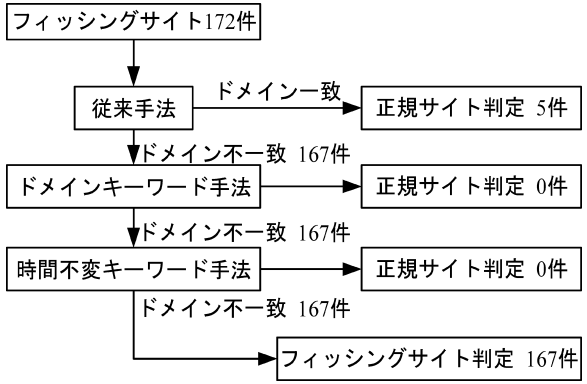

図 8　フィッシングサイトの実験結果
Fig. 8　Experimentation result for phishing sites.

8.1　経験則方式との比較

文献 2) によると，正規サイトをフィッシングと誤検知する率は，基本的コンテンツベース方式では 6%であったが，経験則を利用することにより 1%に減少した．一方，7 章で述べたように，本論文の評価では，基本的コンテンツベース方式の誤検知率は 14.0%，提案方式の誤検率は 7.6%であった．評価サンプルが異なるため効果の厳密な比較はできないが，経験則利用方式の効果は大きいといえる．

しかし，経験則利用方式には，フィッシングサイトの検知率が基本的コンテンツベース方式の 97%から 89%に低下するという問題があった．さらに，フィッシングサイト製作者が経験則を知っていると，それを利用して検知を回避することが可能である．そのため，検知率がさらに大幅に低下すると考えられる．たとえば，3.1 節で説明した経験則のうち ② 以降はきわめて容易に回避できる．経験則 ① についても，安価なドメインを 1 年以上借りてからフィッシングを開始する，あるいは，1 年以上経ったドメインのページを乗っ取ってフィッシングに利用するなどの方法で回避できる．

これに対し，提案方式は，少なくとも 7.2 節の評価実験の範囲では，フィッシングサイトの検知率を低下させていない．また，提案方式を回避するためには，フィッシングページまたは同一ドメインのページが検索エンジンの上位にランクされることが必要であるが，フィッシングサイトの平均寿命は 3.1 日[8] であり，他のウェブページからリンクが張られる

ことが稀であるため，検索エンジンの性質から困難である．

8.2 Web of Trust 方式との比較

Web of Trust 方式の評価について，文献 12) は 2 セットの誤検知率データを記載している．第 1 の誤検知率データは，ページタイトルをキーワードとして正規サイト候補を検索した後，基本的コンテンツベース方式の判定と Web of Trust 方式の判定を各々行い，結果を比較したものである．基本的コンテンツベース方式の誤検知率は 36.1%，Web of Trust 方式の誤検知率は 13.4%であった．Web of Trust 方式の評価サンプルは日本語サイトであり，本論文の評価サンプルは英語サイトであるため，サンプルの性質が大きく異なる．そのため，Web of Trust 方式の改善率と提案方式の改善率を単純に比較することはできないが，Web or Trust 方式に一定の効果が認められる．

第 2 の誤検知率データは，銀行名をキーワードとした場合であり，誤検知率は 0%となっている．しかし，検査対象ページから銀行名を自動的に抽出するアルゴリズムは現時点では確立されていない．

Web of Trust 方式では，検査対象ページに至るリンク元ページを検索できない場合に誤検知が発生する．そのため，リンク元のページを検索するためのキーワード選定アルゴリズムが誤検知率の支配要因となる．一方，提案方式は，検査対象ページと同じドメインのページを検索できない場合に誤検知が発生する．そのため，同一ドメイン検索のためのキーワード選定アルゴリズムが誤検知率の支配要因となる．

このように，Web of Trust 方式と提案方式は独立な性質を持つため，2 つの方式を組み合わせることで誤検知をさらに低減できる可能性がある．たとえば，一方の方式でフィッシング判定となった場合に他の方式で再判定する方法が考えられる．これらの方法を実装，評価することは今後の課題である．

なお，文献 12) の Web of Trust 方式では，4 章の誤検知原因のうち自然言語処理の失敗（4.3 節）および検索エンジンのキャッシュとの相違（4.4 節）を想定していない．文献 12) のキーワード選定アルゴリズムは，ページタイトルおよび銀行名をキーワードとするので，これらの問題が顕在化していないが，将来的に自然言語処理に基づくキーワード選定アルゴリズムを用いる場合には顕在化する可能性がある．すなわち，自然言語処理の失敗による不自然な単語やニュースなどに現れる一時的な単語を選択してしまい，リンク元ページの検索に失敗する可能性がある．この点は，Web of Trust 方式の本質的な問題ではないが，キーワード選定アルゴリズムを改良してゆくうえでは問題となりうる．

提案方式のドメインキーワード手法および時間不変キーワード手法は，自然言語処理の失敗および検索エンジンのキャッシュとの相違に対応している．そこで，Web of Trust 方式のキーワード選定アルゴリズムを改良する際に，これらのキーワード選定手法を参考にできる可能性がある．

8.3 コンテンツ一致方式

コンテンツ一致方式[13] は，正規サイトの検索に失敗したので判定が信頼できないことを検知する．しかし，この方式は，判定が信頼できないことを示すだけであり，正規かフィッシングかの判定精度を向上させる効果はない．そのため，コンテンツ一致方式は単独では効果が少ない．

コンテンツ一致方式と提案方式を組み合わせることで誤検知率を低減させることができる可能性はある．たとえば，提案方式でフィッシング判定となった場合に，コンテンツ一致方式で判定の信頼性を評価し，判定が信頼できない場合には，提案方式を再度実行し，優先度の低かったキーワードを用いて正規サイト候補を再検索するという方法が考えられる．これらの方法を実装，評価することは今後の課題である．

9. ま と め

コンテンツベース方式によるフィッシング検知には，正規サイトをフィッシングサイトであると誤検知することが多いという問題がある．本論文では誤検知の原因を，検索エンジンに載らないページ，複数ドメインで運営されているサイト，自然言語処理の失敗，検索エンジンのキャッシュとの相違の 4 つに分類した．

誤検知の原因に対応した解決手法として，ドメインキーワード手法と，時間不変キーワード手法の 2 つを提案した．ドメインキーワード手法は，検査対象サイトだけでなく周辺ページを含む複数ページを解析することでより広域の特徴を抽出する手法である．時間不変キーワード手法は，閲覧サイトの過去ページを参照することで時間的に安定な特徴を抽出する手法である．

提案手法を実装し，評価実験を行った．結果を**図 9** に示す．正規サイトをフィッシングサイトと誤検知する率は，従来手法が 14.0%であったのに対して，提案手法では 7.6%に減少した．また，フィッシングサイトを正規サイトと誤検知する率は，従来手法も，提案手法も 2.9%のままであり，変化がなかった．以上から提案手法の有効性を検証することができた．

謝辞　本研究は，文部科学省特定領域研究「情報爆発時代に向けた新しい IT 基盤の研究」課題番号 B01-09「ナイーブなユーザのための安全・安心情報生活基盤の研究」（H19-20）の助成を受けて行われた．

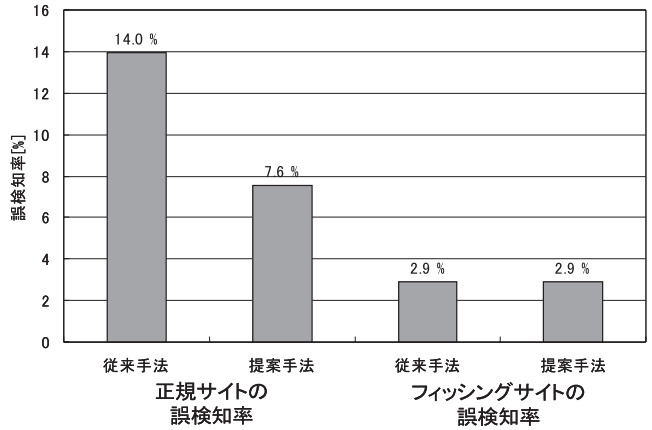

図 9　正規サイトとフィッシングサイトの誤検知率の比較
Fig. 9　Comparison of false detection rates of legal and phishing sites.

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（2008 年 11 月確認）
2) Zhang, Y., Hong, J. and Cranor, L.: 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, No.71, pp.387–392 (2007).
4) 柴田賢介，荒金陽助，塩野入理，金井　敦：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/（2008 年 11 月確認）
6) 中村元彦，寺田真敏，千葉雄司，土居範久：proxy を利用した HTTP リクエスト解析による AntiPhishing システムの提案，2006-CSEC-32, Vol.2006, No.26, pp.13–18 (2006).
7) APWG. http://www.antiphishing.org/（2008 年 11 月確認）
8) Phishing Activity Trends Report for the Month of January, 2008. http://www.antiphishing.org/reports/apwg_report_jan_2008.pdf（2008 年 11 月確認）
9) Dhamija, R., Tygar, J.D. and Hearst, M.: Why Phishing Works, *CHI2006* (2006).
10) Liu, W., Deng, X., Huang, G. and Fu, A.Y.: An Antiphishing Strategy Based on Visual Similarity Assessment, *IEEE Internet Computing*, Vol.10, No.2, pp.58–65 (2006).
11) 長尾　真：岩波講座ソフトウェア科学 15 自然言語処理，岩波文庫，p.421 (1996).
12) 西垣正勝，長谷　巧，原　正憲：Web of Trust の導入によるコンテンツベースフィッシング検知方式の改良，情報処理学会論文誌，Vol.49, No.9, pp.3104–3111 (2008).
13) 中山心太，内海　彰，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の実装と評価，2008-CSEC-40, Vol.2007, No.21, pp.273–278 (2008).
14) Internet Archive. http://www.archive.org/（2008 年 11 月確認）
15) TreeTagger. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/（2008 年 8 月確認）
16) [を] 形態素解析と検索 API と TF-IDF でキーワード抽出. http://chalow.net/2005-10-12-1.html（2008 年 8 月確認）
17) Banks<Financial Services in the Yahoo! Directory. http://dir.yahoo.com/Business_and_Economy/Shopping_and_Services/Financial_Services/Banking/Banks/（2008 年 8 月確認）
18) PhishTank | Join the fight against phishing. http://www.phishtank.com/（2008 年 8 月確認）
19) 中山心太，内海　彰，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の実装と評価，2008-CSEC-40, Vol.2007, No.21, pp.273–278 (2008).

(平成 20 年 12 月 3 日受付)
(平成 21 年　6 月 4 日採録)

中山　心太（正会員）

2007 年電気通信大学電気通信学部電子工学科卒業．2009 年電気通信大学大学院電気通信学研究科人間コミュニケーション学専攻博士前期課程修了．現在，日本電信電話株式会社 NTT 情報流通プラットフォーム研究所に勤務，情報セキュリティの研究に従事．

吉浦　　裕（正会員）

1981 年東京大学理学部情報科学科卒業．日立製作所を経て，2003 年より電気通信大学電気通信学部人間コミュニケーション学科勤務．現在，同教授．自然言語処理，知識処理の研究を経て，現在，情報セキュリティの研究に従事．理学博士．電子情報通信学会，日本セキュリティ・マネジメント学会，システム制御情報学会，人工知能学会，IEEE 各会員．2000 年日立製作所社長技術賞，2005 年情報処理学会論文賞，2005 年システム制御情報学会産業技術賞，2006 年 IEEE International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Best Paper Award 各受賞．