社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－ＤＰＳ－134 (47)
2008－CSEC－ 40 (47)
2008／3／7

# 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

中山 心太[*1] 内海 彰[*2] 吉浦 裕[*3]

[*1]電気通信大学 電気通信学研究科 人間コミュニケーション学専攻

[*2]電気通信大学 電気通信学部 システム工学科

[*3]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

〒182-8585 東京都調布市調布ヶ丘 1-5-1

EMail: [*1]shinta@edu.hc.uec.ac.jp [*2]utsumi@se.uec.ac.jp [*3]yoshiura@hc.uec.ac.jp

**概要** 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，それらの特性を利用したコンテンツベースのフィッシング検知方式が提案されている．しかしこの方式には正規サイトの誤検知率が高いという問題点がある．我々は正規サイトの誤検知を防ぐ手法として，コンテンツ一致方式の構想を提案していた．本稿では未実装であったコンテンツ一致方式を実装，評価した．これにより，正規サイトの誤検知率を低下させる見通しを得ることができた．

**キーワード** フィッシング詐欺，ネットワークセキュリティ，ウェブ

# Implementation and Evaluation of Phishing detection based on features of mimic content.

Shinta Nakayama[*1] Akira Utsumi[*2] Hiroshi yoshiura[*3]

[*1]The Department of Human Communication, The Graduate School of Electro-Communications, The University of Electro-Communications

[*2]The Department of Systems Engineering, The Faculty of Electro-Communications, The University of Electro-Communications

[*3]The Department of Human Communication, The Faculty of Electro-Communications, The University of Electro-Communications

1-5-1,Chofugaoka, chofu-shi 182-8585,Japan

EMail: [*1]shinta@edu.hc.uec.ac.jp [*2]utsumi@se.uec.ac.jp [*3]yoshiura@hc.uec.ac.jp

**Abstract** Most phishing sites are mimics of legal sites. A phishing detection method, called a content-based method, uses this feature of phishing sites. However, the content-based method judges legal sites as phishing with high ratios. Although our earlier paper proposed content-similarity judging to avoid these false positives, the content-similarity judging was not implemented. This paper describes the implementation of the content-similarity judging and the evaluation of its effect on avoiding false detections.

**Keywords** Phishing attack, Network security, web

## 1.はじめに

子供や高齢者などコンピュータリテラシーの低い層のインターネット利用が一般化してきた．これに伴い，低リテラシー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い個人情報を盗み取ることを目的としたウェブサイトを製作し，これによって得られたクレジットカード番号や預金口座の暗証番号，社会保障番号などを悪用し金銭を得る詐欺である．

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の誤検知率が高いという問題がある．

そこで，いくつかの改善手法が提案されている．Yueらによる経験則を用いたCANTINA方式[1]，中山らによるコンテンツの類似性を推し進めた方式[2]，長谷らによるWeb of Trustを用いた方式がある[3]．

本稿では，中山らの方式で未実装であったコンテンツ一致の実装を行い，その評価を行った．

## 2.先行研究

### 2.1 リスト方式

#### 2.1 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，載っていないいサイトを弾く方式である[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．

#### 2.1.2 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弾く方法である[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，登録されていく．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．

### 2.2 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日と非常に短い[8]．そのため，ウェブ存続期間，ドメインの登録日時，DNSの逆引きが可能かどうか，GoogleのPageRank等を調べることで，フィッシングサイトか否かの判定を行うことができる．

### 2.3 コンテンツの特性に基づいた方式

フィッシングサイトはユーザーを騙すために特定の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は酷似しているといえる．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．そのため，コンテンツの類似性を利用した手法が提案されている．

#### 2.3.1 視覚的類似性に基づいた方式

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手法がLiuらにより提案されている[9]．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，ユーザサイドで判定することはできない．

#### 2.3.2 テキストの類似性に基づいた方式

フィッシングサイトのテキストの類似性を利用し，フィッシング検知を行う手法(テキストコンテンツベース方式)である[1][2]．フィッシングサイトに含まれる語句には，正規サイトで利用しているものが多く含まれている．自然言語処理でそれらの語句を抽出し，その語句をキーワードとして検索を行うことで正規サイトを発見することができる．そしてフィッシングサイトと正規サイトを比較して，フィッシングサイト判定を行う[図1]．

$$w(t,d)=tf(t,d)\cdot idf(t)$$

$$idf(t)=\log\left(\frac{S}{df(t)}\right)$$

式1:TF-IDF法の計算式

なお，先行研究では自然言語処理にTF-IDF法[11]が用いられている．TF-IDF法は文章中の単語の重みを計算する手法で，式1で表される．文章d中の単語tの重みwを，tの出現回数tfと，他の文章にtがどれほど現れているか，という単語の稀少性idfの積によって定義する．Sはサンプル文章の総数，df(t)はサンプル母集団中に単語tが含まれる文章の数を表す．

テキストコンテンツベース方式は検索エンジンを利用する．検索エンジンは，古くから存在し，頻繁に更新され，他のサイトからリンクが多く張られているサイトを高く評価するようにできている[10]．逆を言えば，フィッシングサイトのような新たにできた他からリンクが張られていないサイトの評価を低く扱う．そのため，フィッシングサイトに正規サイトに同じキーワードが含まれていたとしても，そのキーワードで検索すると，正規サイトのほうが上位に来る．これにより，検索エンジンをある種のホワイトリストして利用することができる．しかしこの手法には，フィッシングサイトの検知率は高いが，正規サイトの誤検知率が高いという問題点がある．

## 3 テキストコンテンツベース方式の課題と改善手法

### 3.1 正規サイトの誤検知の発生理由

テキストコンテンツベース方式は，正規サイトの誤検知率が高いという問題がある．その原因は中山[2]，長谷[3]らが分析し，以下の4点にまとめた．

#### 3.1.1 TF-IDF法による特徴語抽出の失敗

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．たとえば「goo+gle」で検索を行ったとしても，googleのページは出てこない．TF-IDF法でそのような単語を特徴語として検索キーワードに選んでしまった場合，自身を検索することが

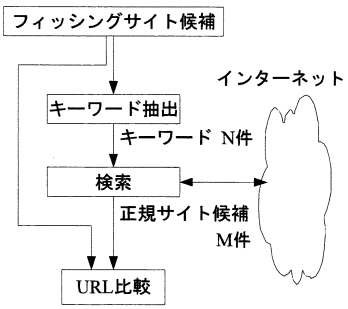


図1:テキストコンテンツベース方式の流れ

できず，フィッシングサイト扱いされる．

### 3.1.2 検索エンジンのキャッシュとの差異

検索エンジンはウェブサイトを定期的に巡回して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，検索エンジンが保持しているウェブサイトとが異なる場合がある．その際に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特徴語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．

### 3.1.3 検索エンジンに乗らないサイト

robots.txtやmetaタグを利用することで，ウェブページは検索エンジンに捕捉されなくなる．そのため，自分自身を検索することができず，フィッシングサイト扱いされる．

### 3.1.4 複数ドメインで運営されているサイト

複数のドメインで運営されているウェブサイトが存在する．たとえば顧客情報を扱うためのサーバを別ドメインで運営していたり，親会社と子会社で別のドメインであったりする．そのようなケースの場合，使われている言葉は似通っているが，ドメインが異なるというケースになってしまい，フィッシングサイトと判断されてしまう．

## 3.2 改善手法

3.1の問題に対応するために，Yue,中山，長谷らはそれぞれ独自のアプローチで改善を行っている．

### 3.2.1 経験則による改善 [1]

YueらのCANTINA方式では，テキスト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 ドメイン登録からの日数が短いか

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る．したがって89%よりも大幅に検知率が下がることが予想さる．そのため，経験則の併用はフィッシング詐欺の検知手法として利用することは難しいと考えられる．

### 3.2.2 コンテンツ一致による改善 [2]

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

| | コンテンツが一致 | コンテンツが不一致 |
|---|---|---|
| URLが一致 | ①正規サイト | ②正規サイト(別ページ) |
| URLが不一致 | ③フィッシングサイト | ④検索失敗 |

中山らの原論文ではフィッシングサイトの判定において，URL判定に加えて，コンテンツ一致判定を行うことを提案していた[表1]．これによりURLが不一致の場合でも，単純にフィッシングサイトと判断しないので，正規サイトの誤検知率を下げることができる．しかし，コンテンツ一致判定は詳細検討を行っておらず，未実装であった．

### 3.2.3Web of Trustによる改善 [3]

長谷らはWeb of Trustの考えを導入した．Web of Trustは『自分の知らない相手が信用できるか否か判断したい場合に，その相手が多数の人から信用されている場合や，自分の信用している他人がその相手を信用している場合には，その相手は信用できると判断する』という考えに基づいている．

検索された正規サイト候補のウェブサイトは十分に信用できると考え，そのウェブサイトからリンクを辿り，フィッシングサイト候補のページにたどり着ければ，そのページは信用できるとし，正規サイトであると判断する．これにより誤検知を防ぐことができる．

## 4 コンテンツ一致方式

中山らはコンテンツ一致方式を提案していたが，詳細な検討はされていなかった．本稿ではコンテンツ一致方式の検討を行い，コサイン類似度によるコンテンツ一致方式の実装を行った．

### 4.1 処理の流れ

URL一致と，コンテンツ一致の二つの軸を用いて，表1を詳細に分析すると次のようになる．

① URLが一致していて，コンテンツが一致している場合．
自分自身を検索できたということなので正規サイトであると判断する．

② URLが一致していて，コンテンツが不一致の場合．
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 URLが不一致で，コンテンツが一致している場合．
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，ユーザーを騙すためにほぼ同一であると考えられる．そのためフィッシングサイトであると判断する．

④ URLが不一致で，コンテンツが不一致の場合．
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と判断できる．

これにより，検索失敗だと判断できるため，検索キーワードの変更や，Web of Trust方式の導入により，誤検知を減らすことができる．

以上から，コンテンツ一致方式のフローチャートは図2のようになる．①と②はどちらも正規サイトと判断するため，URLが一致していれば，正規サイトと判断するようにまとめられる．URLが一致してなかった場合，コンテンツの一致判定を行うことで，本当にフィッシングサイトだったのか，検索失敗なのかを判断することができ，これにより正規サイトの誤検知率を下げることができる．

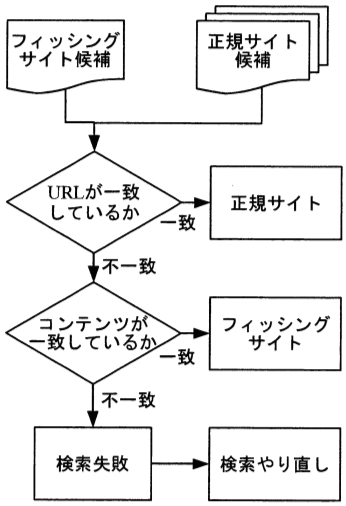


図2:コンテンツ一致方式のフローチャート

### 4.2 実装

コンテンツ一致の判定はコサイン類似度[16]を用いて実装を行った．コサイン類似度は二つの多次元ベクトルが与えられていた時に，その角度(コサイン)を求めることで，ベクトル間の類似度を算出する手法である．

二つの文章 $d_1$，$d_2$ からそれぞれ単語 t の特徴度ベクトル $v_1$，$v_2$ を作成し，そのベクトルのコサインを求めることで類似度とする[式2]．また，特徴度ベクトルの要素 w(t, d) は文章 d 中の単語 t の特徴度を表す．

なお，$d_1$，$d_2$ で利用されている言葉が完全一致していれば1になり，利用している言葉がまったく異なれば，特徴度ベクトルが直交するので0になる．また，$v_1$，$v_2$ はすべての要素が0以上であるので，コサイン類似度の変域は[0, 1]になる．

コサイン類似度の計算例を次に示す．特徴度 w(t, d) は式1の TF-IDF 法を利用する．

$$\begin{cases} v_1 = v(d_1) = [w(t_1, d_1), w(t_2, d_1), w(t_3, d_1), \ldots, w(t_m, d_1)] \\ v_2 = v(d_2) = [w(t_1, d_2), w(t_2, d_2), w(t_3, d_2), \ldots, w(t_m, d_2)] \end{cases}$$

$$\text{類似度} = \cos(v_1, v_2) = \frac{v_1 \cdot v_2}{|v_1||v_2|}$$

$$= \frac{\sum_{k=1}^{m} (w(t_k, d_1) \cdot w(t_k, d_2))}{\sqrt{\sum_{k=1}^{m} (w(t_k, d_1))^2} \sqrt{\sum_{k=1}^{m} w((t_k, d_2))^2}}$$

式2:コサイン類似度の計算式

次の文章 $d_1$，$d_2$ があるとする．この文章の名詞 t の特徴度 idf(t) と出現回数 tf(t, d) を求めると表2のようになる．

$d_1$：今日の天気は晴れです．明日は雨でしょう．

$d_2$：今日の天気は雨です．明日も雨です．

表2:名詞 t の idf と tf

| 単語 | 今日 | 天気 | 晴れ | 明日 | 雨 |
|---|---|---|---|---|---|
| idf(t) | 9.65 | 15.14 | 14.89 | 11.00 | 10.01 |
| tf(t, $d_1$) | 1 | 1 | 1 | 1 | 1 |
| tf(t, $d_2$) | 1 | 1 | 0 | 1 | 2 |

以上から特徴度ベクトル v($d_1$)，v($d_2$) を求めると次のようになる．

$$\begin{cases} v(d_1) = [w(\text{今日}, d_1), w(\text{天気}, d_1), w(\text{晴れ}, d_1), w(\text{明日}, d_1), w(\text{雨}, d_1)] \\ v(d_2) = [w(\text{今日}, d_2), w(\text{天気}, d_2), w(\text{晴れ}, d_2), w(\text{明日}, d_2), w(\text{雨}, d_2)] \end{cases}$$

$$v(d_1) = \begin{bmatrix} 1\times9.65 \\ 1\times15.14 \\ 1\times14.89 \\ 1\times11.00 \\ 1\times10.01 \end{bmatrix}^T \quad v(d_2) = \begin{bmatrix} 1\times9.65 \\ 1\times15.14 \\ 0\times14.89 \\ 1\times11.00 \\ 2\times10.01 \end{bmatrix}^T$$

$$v(d_1) = \begin{bmatrix} 9.65 \\ 15.14 \\ 14.89 \\ 11.00 \\ 10.01 \end{bmatrix}^T \quad v(d_2) = \begin{bmatrix} 9.65 \\ 15.14 \\ 0 \\ 11.00 \\ 20.02 \end{bmatrix}^T$$

特徴度ベクトル v($d_1$)，v($d_2$) から，コサイン類似度を求めると次のようになる．

$$\cos(v(d_1), v(d_2)) = \frac{v(d_1) \cdot v(d_2)}{|v(d_1)||v(d_2)|} = \frac{643.7}{\sqrt{765.2} \cdot \sqrt{844.1}} = 0.80$$

コサイン類似度が 0.80 であるため，文章 $d_1$ と $d_2$ はよく類似しているといえる．

## 5 英語のウェブサイトへの対応

フィッシング対策協議会[13]のレポート[14]によると，日本語のフィッシングサイトは月に数十件しか出現していない．しかし，英語のフィッシングサイトは一日に約千件も現れている．そのため，英語のフィッシングサイトへの対応を行った．

TF-IDF 法で単語の重みを評価するためには，文章を単語単位に分割する必要がある．日本語であれば形態素解析などが必要になるが，英語は分かち書きがされているので，そのよ

うな処理は必要としない．しかしすべての単語を評価していては効率が悪いため，TreeTagger[15]を利用して英文の品詞情報を取得し，名詞のみを解析対象とした．

TreeTaggerを利用して，"This paper is a good paper."という文章を品詞解析すると表3のようになる．paperがNN(noun)なので，名詞であることが分かる．

表3:TreeTaggerによる品詞解析結果

| 単語 | 品詞 | 原形 |
|---|---|---|
| This | DT | this |
| paper | NN | paper |
| is | VBZ | be |
| a | DT | a |
| good | JJ | good |
| paper | NN | paper |

## 6 評価

北米の銀行のウェブサイトのトップページ47件を対象に，正規サイトの検知実験を行った．実験結果は表4のようになった．検索結果0件とは，検索キーワードの選定に失敗し，ウェブサイトが一件も検索できなかったことを意味する．

実験に用いたウェブサイト47件中10件の誤検知が発生した．そのうち2件はキーワード選定に失敗して，検索ができず，正規サイト候補が得られなかったものである．

正規サイト候補が得られたが，URLが不一致であったもののコサイン類似度を表5に示す．なお，コサイン類似度は正規サイト候補の中で，最もコサイン類似度が高かったものである．

表4:実験結果

| | 件数 |
|---|---|
| **URLが一致** | 37件 |
| **URLが不一致** | 8件 |
| **検索結果0件** | 2件 |

ここでは，コサイン類似度が0.7を超えると，コンテンツが一致したと判断する．このようにすると，case1～4はURLが不一致で，コンテンツが一致しているといえる．そのため，表1の③にあたり，フィッシングサイトだと判断される．しかし，case1～4はブラウザからアクセスすると，リダイレクトされて，別のURLに転送される．case1～4の正規サイト候補を調べると，リダイレクト先のURLが含まれていた．実験に用いたプログラムでは，HTTPのGET時にステータスコードが301 Moved Permanentlyであったときに，自動的にリダイレクト先を取得するようになっていた．そのため，URLはリダイレクト前，コンテンツはリダイレクト後という状態になってしまった．そして，自分自身が検索できていたのにもかかわらず，フィッシングサイトであると誤検知していた．そのため，リダイレクト先のURLを現在のURLであるとみなすようにプログラムを修正することで，case1～4は正規サイトであると判断することができる．そのため，これらは従来との差分にはならない．

case5～8はURLが不一致で，コンテンツも不一致であるため表1の④にあたり，検索失敗だと判断される．そのため，キーワード選定をやり直し，正規サイトを選出しなおすこと

表5　検知失敗したケースの分析

| case | URL | コサイン類似度の最大値 | 原因 |
|---|---|---|---|
| 1 | www.bankofbotetourt.com | 1.00 | www.bankofbotetourtonline.comにリダイレクト。 類似ドメインの事前確保？ |
| 2 | www.bankamerica.com | 1.00 | www.bankofamerica.comにリダイレクト。 類似ドメインの事前確保？ |
| 3 | www.bankone.com | 1.00 | www.chase.comにリダイレクト。 銀行合併によるURL変更 |
| 4 | www.asbnj.com | 0.75 | www.americansavingsnj2.comにリダイレクト。短縮URL？ |
| 5 | www.bofm.com | 0.56 | 銀行とは関係ない一般名詞をキーワードとして拾ってしまった。 |
| 6 | www.bancopopular.com | 0.54 | トップページにテキストはpopularの一語しかなかった。 |
| 7 | www.audubonsavings.com | 0.43 | 画像が多く、キーワードとなるテキストを拾えなかった。 |
| 8 | www.4thebank.com | 0.01 | 画像が多く一般名詞のキーワードしか拾えなかった |

表6 http://www.audubonsavings.com/　の検索キーワードとコサイン類似度

| **検索キーワード** | Parish,Natchitoches,Montgomery,Louisiana,Grant |
|---|---|
| **Googleで検索された正規サイト候補のURL** | **コサイン類似度** |
| http://local.yahoo.com/NJ/Pine+Hill/Entertainment+Arts/Entertainment+Venues | 0.39 |
| http://local.yahoo.com/NJ/Pine+Hill/Food+Dining/Restaurants/Ice+Cream+Yogurt | 0.36 |
| http://www.sjsca.org/directors.cfm | 0.35 |
| http://www.pinehillboronj.com/default.htm | 0.15 |
| http://www.switchboard.com/swbd.main/dir/results.htm?MEM=46&KW=Bars&LO=City+of+mount+laurel%2C+NJ | 0.21 |
| http://www.scnes.com/erial-nj-dishnetwork-tv-deals.html | 0.30 |
| http://www.retirenet.com/location/communities/149-new-jersey/1-active-lifestyles?city=08081-sicklerville | 0.18 |
| http://dsl.newtechnologytv.com/erial-nj-broad-band-dsl.html | 0.31 |
| http://www.local-worship.com/Laurel_Springs-NJ.html | 0.43 |
| http://www.wesleyberryflowers.com/flowers_erial_NJ.html | 0.38 |

で，誤検知を改善できる可能性がある．従来手法であればこれらはすべてフィッシングサイトだと判断されていた．

表6にcase7の検索キーワードとコサイン類似度出力結果を示す．正規サイト候補のURLの中に同じドメインのサイトが無いため，URL一致に失敗していることが分かる．また，コサイン類似度は，最高でも0.43と低いので，正規サイト候補中に正規サイトがないことが分かる．そのため，キーワード選定をやり直し，正規サイト候補を選出しなおす，Web of Trust方式を導入するなどで，誤検知を改善できる可能性がある．また，case5,6,8も同様である．

## 7 まとめ

コンテンツ一致方式のフィッシングサイト検知手法の実装，評価を行った．

正規サイトの検知実験を行った．従来手法であれば47件中6件を誤検知していた．しかし，コンテンツ一致の手法を用いることで，そのうち4件を正規サイト候補の検索に失敗したと判断できた．これにより，検索失敗と分かったら，検索をやり直したり，Web of Trust方式を導入することで，誤検知率を低下させることができることが分かった．そのため，コンテンツ一致方式により，正規サイトの誤検知率を低下させることができるという見通しを得た．

今後の課題は次のとおりである．

- 検索失敗時の処理
  検索失敗と判断したときにどのような処理をするか，どのように検索をやり直すのかの検討をする。
- コサイン類似度の閾値の決定
  今回の評価実験では，誤検出が発生したケースが少なく，コサイン類似度の閾値を設定することができなかった．そのため，より多くの実験を行い，コンテンツ一致の閾値を決定する．

## 謝辞

本研究は文部科学省科学研究補助金「特定領域研究」「情報爆発時代に向けたIT基盤技術の研究」（平成１９―２０年度）の支援を受けている．

また，フィッシング対策協議会様に最新のフィッシングサイトの実態，対策の動向についてご教示いただいた．ここに深く謝意を表す．

## 参考文献

[1]Yue Zhang, Jason Hong, Lorrie Cranor, ” CANTINA: A Content-Based Approach toDetecting Phishing Web Sites”, WWW2007, 2007.

[2]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,Vol2007, No71,pp387-392,2007.

[3]長谷巧,原正憲,西垣正勝,"Web of Trust の導入によるコンテンツベースフィッシング対策の改良",CSS2007pp.525-530,2007

[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

[5]"RBL.JP",http://www.rbl.jp/ （2008年2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

[7]"APWG", http://www.antiphishing.org/ （2008年2月確認）

[8]"Phishing Activity Trends Report for the Month of November, 2007" http://www.antiphishing.org/reports/apwg_report_nov_2007.pdf （2008年2月確認）

[9]Liu Wenyin, Guanglin Huang, Liu Xiaoyue ,Zhang Min ,Xiaotie Deng, "Detection of Phishing Webpages based on Visual Similarity",Proc.of WWW2005, pp1060 - 1061

[10]"Google の秘密 - PageRank 徹底解説", http://www.kusastro.kyoto-u.ac.jp/~baba/wais/pagerank.html （2008年2月確認）

[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 pp118-130

[12]"[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2月確認）

[13]"AntiPhishingJapan フィッシング対策協議会",http://www.antiphishing.jp/ （2008年2月確認）

[14]"フィッシング対策協議会4半期レポート 2007年10-12月期",http://www.antiphishing.jp/report/200802-case-110.pdf （2008年2月確認）

[15]"TreeTagger", http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/ （2008年2月確認）

[16]北研二，津田和彦，獅々掘正幹,"情報検索アルゴリズム", 共立出版株式会社 pp61