Vol.2019-IOT-47 No.3
Vol.2019-SPT-35 No.3
2019/9/19

URL文字列を用いたフィッシングサイト検知手法の提案（2019年8月19日版）

片山 皐佑[1,a)]　今泉 貴史[2,b)]

概要：インターネット利用者の増加に伴い，サイバー犯罪が増加，高度化している．サイバー犯罪の1つであるフィッシング攻撃もその例外ではない．フィッシング攻撃に使用される，フィッシングサイトの生存期間は非常に短く，ブラックリストの作成による対策が難しい．そのためフィッシングサイトの動的検知が望まれる．本研究では，フィッシングサイトのURL文字列の特徴を用いることで，短時間での動的検出を行う手法を提案する．判別の要素として，URL文字列中の文字の出現頻度，出現確率，ドメインの長さを用いる．それぞれの要素が正規のWebサイトからどれだけ離れているかを計算し，それらの値を基にフィッシングサイトかどうか判別を行う．

Proposal of phishing site detection method using URL string (version 2019/8/19)

1. はじめに

インターネットの普及に伴い，オンラインバンクやECサイトを利用する顧客が増加してきている．実際のサイトを模倣して作ったサイトのリンクをメールなどでユーザに送り，クレジットカードの番号などの個人情報を盗み出すフィッシング攻撃も増加してきている．フィッシング攻撃は，ユーザがメールなどで送られてきたフィッシングサイトのリンクを開き，正規のサイトだと勘違いして個人情報を入力することにより行われる．

現在フィッシングサイトの検知手法としてWebサイトのコンテンツを調べることによって検知を行なう手法やgoogleなどの検索エンジンを用いて検知を行なう手法が多く提案されている．前者は検知に非常に長い時間がかかり，後者は新規に作成されたWebサイトの誤検知が起こる．

本研究では，フィッシングサイトの検知にURL文字列を用いることで，短時間かつ正規サイトの誤検知を少なくすることを目的とする．単純な文字列の処理のみを行い判別を行なうため，非常に短い時間い時間で処理を行える．そのため，ユーザがフィッシングサイトへのアクセスを行う前に検知を行い，フィッシングサイトへのアクセスを防止することを目的とする．単純な文字列処理のみを行なうため，フィッシングサイトの検知にかかる時間が非常に短くなる．

以下，2章では，フィッシングについて説明する．3章では，提案手法について説明し，4章では提案手法を用いた実験を行う．5章では，実験についての考察を行なう．最後に6章でまとめと今後の課題について述べる．

2. フィッシング

2.1 フィッシングサイトの特徴

フィッシングサイトは，ユーザを騙し個人情報を盗むために作成されたサイトである．そのため，フィッシングサイトはユーザに勘違いさせる正規サイトを模倣して作られている．ゆえに，フィッシングサイトの中には正規サイトで使われている画像や正規サイトのブランド名などが含まれている．また，フィッシングサイトはユーザの個人情報を盗む目的のみで作られていることが多く，サイト内に含まれるリンクの数が少ないことがある．さらに，APWGの調査によると，フィッシングサイトの生存期間は平均4

[1] 千葉大学大学院融合理工学府
Graduate School of Science and Engineering, Chiba University
[2] 千葉大学統合情報センター
Institute of Management and Information Technologies, Chiba University
[a)] kosukekatayama_0518@chiba-u.jp
[b)] imaizumi_takashi@faculty.chiba-u.jp

表 1　模倣して作られたフィッシングの URL
Table 1　Phishing URL.

実際のドメイン	模倣されたドメイン
apple.com	apple-com, applecom, apple.con appile.com, app-le.com
paypal.com	pavpal.com, puaypal.com pauypal.com paypal-com, paypai.com
wells.fargo.com	wellsf.argo.com

日，最長でも 30 日と短い [3]．それにも関わらず攻撃者はフィッシングサイトがより正当なものに見えるように証明書を取得している．F5 Network Inc の 2018 年 9 月から 10 月にかけての調査では，収集したフィッシングサイトのうち，93%が https を利用していたと報告されている [4]．フィッシングサイトを生成する際の URL に関しても例外でない．ランダムな文字列の URL を自動生成する方法から URL も視覚的に騙すためのものに変化してきている．Wandara の調査で挙げられている，実在するサイトを模倣して作られた URL の一部を表 1 に示す．このように一見本物のドメインと区別のつかないようなものがある．

2.2　フィッシング対策手法

2.2.1　ブラックリスト方式

予めブラックリストに追加しているページへのアクセスをブロックする手法．ユーザからの報告や特定のアルゴリズムによって新しいページをブラックリストに追加する．正規サイトの誤検知が少なく，短時間での検知が可能である．一方，ブラックリストに登録されていない場合，検知漏れが発生する．また，ブラックリストの管理維持にかかるコストが大きい．

2.2.2　ホワイトリスト方式

正規サイトをリスト化し，ホワイトリストに載っていないサイトを弾く手法．フィッシングサイトの検知漏れが無く，短時間で検知が可能である．一方，すべての正規サイトを網羅したホワイトリストの作成は難しく，有名でないサイトや新しく作られた Web サイトなどがホワイトリストへの登録漏れにより，フィッシングサイトとして扱われることがある．

2.2.3　ヒューリスティック

フィッシングサイトの特徴を基に動的解析を行い検知する手法．正確に分類できる訳では無いが，アルゴリズムによって高い確率で正解を期待することができる．手法の例を以下に述べる．

Jun Ho Huh らは，検索エンジンを用いた判別アルゴリズムを提案した [1]．この手法では，検索エンジンの PageRank 等を調べることで，フィッシングサイトであるかどうかの判別を行う．yahoo の検索エンジンを用いた際に，98%を超える精度であった．しかし，個人サーバや新しく作られた Web サイトをフィッシングサイトと判別してしまう可能性がある．

Yue Zhang らは，TF-IDF を用い，ページ内のキーワードを抽出し，判別を行う手法を提案した [2]．6%の偽陽性で 97%のフィッシングサイトを検出できるが，検知を行うのにかかる時間が長い．また，ブランド名が一般名詞からのみで構成されている場合，フィッシングサイトとして判別される可能性が高い．

表 2　各要素の平均と標準偏差
Table 2　Parameters

パラメータ	平均	標準偏差
階層数	μ_{hie}	σ_{hie}
出現頻度値	μ_{prob}	σ_{prob}
最長ドメイン長	μ_{len}	σ_{len}
最小出現確率	μ_{app}	σ_{app}
数字の塊の数	μ_{num}	σ_{num}
ハイフンの数	μ_{hyp}	σ_{hyp}

3.　URL 文字列を用いたフィッシングサイト検知手法

本研究では，フィッシングサイトを検知する際に用いる要素として，URL の文字列のみを用いることで判別を行う．ランダムに生成された URL 文字列のフィッシングサイトだけでなく，正規サイトを模倣して作られた URL 文字列のフィッシングサイトの検知を目標とする．また，新規に作成された正規サイトや検索エンジンにあまり引っかからないような小規模なサイトの誤検知を防ぐことも目指す．

単純な文字列処理のみを行うため，非常に短い時間でフィッシングサイトの検知を行なうことが可能である．正規サイトの URL から特徴を抽出し，未知のフィッシングサイトを検出する．そのため，フィッシングサイトのデータセット無しで判別を行なうことができる．

本研究で判別に用いる要素は，以下のものである．

- ドメインの階層数
- 文字の出現頻度
- ドメインの長さ
- 文字の出現確率
- 数字の出現数
- ハイフンの出現数

これらの要素から個別にフィッシングサイトであるかどうかを判別する危険度を求め，求めた危険度の総和が閾値を超えた場合，フィッシングサイトと判別する．また，これらの要素の平均値と標準偏差を表 2 のように表す．危険度は 0 から 3 の 4 段階で示す．各要素の危険度について表 3 に示す．

次にそれぞれの要素と危険度の決め方について説明していく．

表 3　危険度とパラメータの値
Table 3　A degree of risk and parameters.

危険度	階層数	出現頻度値	最長ドメイン長	最小出現確率	数字の塊の数	ハイフンの数
0	1	$\mu_{prob}+\sigma_{prob}$ 未満	$\mu_{len}+\sigma_{len}$ 未満	$\mu_{app}-\sigma_{app}$ より大きい	0	0
1	2	$\mu_{prob}+\sigma_{prob}$ 以上 $\mu_{prob}+2\sigma_{prob}$ 未満	$\mu_{len}+\sigma_{len}$ 以上 $\mu_{len}+2\sigma_{len}$ 未満	$\mu_{app}-\sigma_{app}$ 以下 $\mu_{app}-2\sigma_{app}$ より大きい	1	1
2	3	$\mu_{prob}+2\sigma_{prob}$ 以上 $\mu_{prob}+3\sigma_{prob}$ 未満	$\mu_{len}+2\sigma_{len}$ 以上 $\mu_{len}+3\sigma_{len}$ 未満	$\mu_{app}-2\sigma_{app}$ 以下 $\mu_{app}-3\sigma_{app}$ より大きい	2	2
3	4 以上	$\mu_{prob}+3\sigma_{prob}$ 以上	$\mu_{len}+3\sigma_{len}$ 以上	$\mu_{app}-3\sigma_{app}$ 以下	3 以上	3 以上

表 4　階層数の例
Table 4　An example of hierarchy.

ドメイン	階層数
example.jp	1
example.co.jp	1
abc.example.jp	2

3.1　ドメインの階層

正規サイトのドメインは，少ない階層のドメインで構成されていることが多い．しかし，一部の ccTLD（国別コードトップレベルドメイン）を使用すると，予め決められている属性型ドメインや地域型ドメインを使うことができる．属性型ドメインであれば「example.co.jp」，地域型ドメインであれば「example.tokyo.jp」のようになる．このように ccTLD に紐づけられたセカンドレベルドメインを使用した場合，セカンドレベルドメインに汎用ドメインを使用した場合と比べて階層が多くなる．そこで，本研究でドメインの階層を扱う場合，任意の文字列を設定できる汎用ドメイン部分の階層を使用して判別を行い，以降の階層数は前述の階層の数を表す．階層数の例について表 4 に載せる．

正規サイトの階層数の多くが 1 である．そのため階層数が 1 増えるごとに危険度を 1 増加させる．

3.2　文字の出現頻度

正規サイトでは，汎用ドメイン部分にブランド名や意味のある単語を使用することが多い．そのため，文字の出現に偏りが生じると考えられる．しかし，URL の文字列は非常に短いため，1 文字単位で出現頻度を求めることが効率的で無い．そこで本研究では，母音（a，i，u，e，o）のグループと出現頻度順に子音を 5 グループに分割して，文字のグループの出現頻度を基に判別を行う．ある URL における文字を分割したグループをそれぞれ g_1，g_2，g_3，g_4，g_5 としたとき，それぞれのグループの出現確率を $P(g_1)$，$P(g_2)$，$P(g_3)$，$P(g_4)$，$P(g_5)$ とする．それぞれのグループの平均の出現確率を $\overline{P(g_1)}$，$\overline{P(g_2)}$，$\overline{P(g_3)}$，$\overline{P(g_4)}$，$\overline{P(g_5)}$ とする．各グループの出現確率と各グループの平均の出現確率の二乗差を以下の式のように求め，その値を判別に用いる．

$$V_{prob}=\sum_{i=1}^{5}(P(g_i)-\overline{P(g_i)})^2$$

V_{prob} を出現頻度値とする．

文字列がランダムに決定された場合，出現頻度値は大きくなると考えられる．そのため出現頻度値が大きくなるにつれて危険度を増加させる．危険度の増加の基準として出現頻度値の平均と標準偏差を利用する．

3.3　ドメインの長さ

フィッシングサイトの URL では，正規サイトの文字列を左側に持ってくることが度々ある．しかし，フィッシングサイトは正規サイトと同一のドメイン名を使用することはできないため，正規サイトドメインの後ろに「-」などを用いて文字列を追加する傾向がある．その結果，最下位ドメインの長さが長くなる．各レベルのドメインのうち最長のドメイン長を基に判別を行う．

最長のドメインの長さが大きくなるにつれて危険度を増加させる．危険度の増加の基準に正規サイトの最長ドメインの平均と標準偏差を利用する．

3.4　文字の出現確率

正規サイトの URL 文字列は，汎用ドメイン部分にブランド名や意味のある単語を使用することが多いため，ある文字の後ろに来やすい文字がある．一方，フィッシングサイトの URL 文字列では，ブランド名の一部を視覚的に欺くための文字に置き換えることがあり，その場合，本来であれば出現する確率が低い文字が出てくることになる．たとえば「example.com」に対して「examp1e.com」のように作られることがある．この場合，「p」の後に「1」が来る確率 $P(1|p)$ と「1」の後に「e」が来る確率 $P(e|1)$ が低いと考えられる．そこで本研究では，ドメイン中の文字列 x の中で最小となる $P(x_{i-1}|x_{i-2})*P(x_i|x_{i-1})$ の値を用いて判別を行う．また，この値を最小出現確率とする．

最小出現確率は，値が小さくなるにつれてフィッシングサイトである確率が上がると考えられるので，値が小さくなるにつれて危険度を増加させる．危険度の増加の基準には，正規サイトにおける最小出現確率の平均と標準偏差を利用する．

3.5 数字の出現数

正規サイトで数字を用いる場合，ドメインの始めや終わりに現れることが多い．また，数字は塊になって出現することが多い．フィッシングサイトでは視覚的に欺くため，「l」の代わりに「1」，「o」の代わりに「0」を用いることがある．また，ランダムに生成した URL では複数の数字がバラバラに出現することがある．そのため，数字の塊がいくつ出現したかを判別の要素に用いる．

数字が使われる正規サイトは多くはない．そのため，数字の塊の出現回数が 0 回のときの危険度を 0 とし，出現回数が増えるにつれて危険度を 1 ずつ増加させる．

3.6 ハイフンの出現数

フィッシングサイトでは，ブランド名の間や実際に「.」がある部分にハイフンを挿入することがある．そのため，ハイフンの数が多くなることがある．そのためハイフンの数を判別の要素に用いる．

正規サイトの多くはハイフンを使用していない．そのためハイフンの出現回数が 0 回のときの危険度を 0 とし，出現回数が増えるにつれて危険度を 1 ずつ増加させる．

4. 実験

前章では，フィッシングサイトの判別に用いる URL 文字列の要素について述べた．本章では，その要素を用いて実験を行い，判別の精度を述べる．述べた要素を基に，フィッシングサイトであるかどうかの判別を行う．各要素において，0 から 3 の 4 段階で危険度を示し，各要素の危険度の総和により判別を行う．本実験では，正規サイトのドメインを Statvoo[5] がランク付けした Top-1million-sites を利用する．取得した正規サイト 100 万件のうち，10 万件を 10 分割して，学習データとする．残りの 90 万件をテスト用のデータとする．フィッシングサイトとして，PhishTank[6] から取得した URL のうち，ドメインが重複しなかった 17256 件を利用する．取得したフィッシングサイトは，すべてテスト用のデータとする．

危険度の決定に要素の平均と標準偏差を使用するものがある．そのため，学習データとして用意した 1 万件の正規サイトを用いて，出現頻度値，最長のドメイン長，最小出現確率の平均と標準偏差を求める．求めた平均と標準偏差から，前述の表 3 に従い危険度の値を決定する．

次にテストデータを用いて判別を行なう．テストデータから汎用ドメイン部分を取り出す際，セカンドレベルドメインが ccTLD に紐づくものなのか判別する必要がある．その判別は，mozilla の TLD List[7] を基に行なう．取り出した汎用ドメイン部分から，判別に用いる各要素の値を計算し，危険度を算出する．算出された危険度の総和を計算し，閾値よりも大きい場合に，フィッシングサイトと判別する．

表 5 フィッシングサイトの検知率
Table 5 Detection rate

危険度の閾値	TP Rate	FP Rate	TN Rate	FN Rate
1	83.7%	58.1%	41.9%	16.3%
2	64.4%	37.9%	62.1%	35.6%
3	49.9%	27.1%	72.9%	50.1%
4	32.3%	14.6%	85.4%	67.7%
5	22.3%	11.3%	88.7%	77.7%
6	16.3%	10.5%	89.5%	83.7%

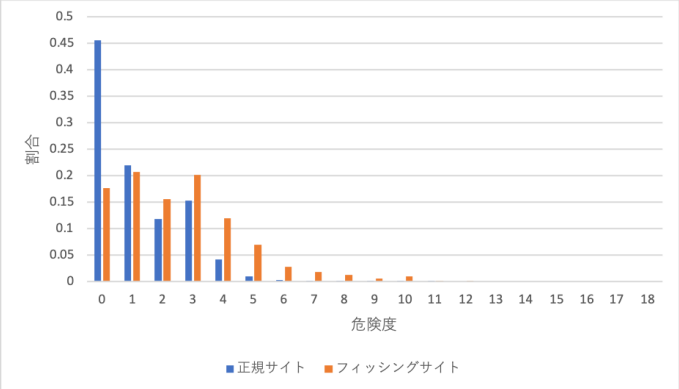

図 1 正規サイトとフィッシングサイトの危険度の分布
Fig. 1 Distribution of degree of risk

判別において，フィッシングサイトと判別することを Positive，正規サイトと判別することを Negative という．ここでフィッシングサイトをフィッシングサイトと正しく判別することを True Positive（TP）といい，正規サイトをフィッシングサイトと誤って判別することを False Positive（FP）という．また，正規サイトを正規サイトと正しく判別することを True Negative（TN）といい，フィッシングサイトを正規サイトと誤って判別することを Fales Negative（FN）という．危険度の閾値を 1 から 6 まで変更して，フィッシングサイトの判別を行った．各学習データを基に行った結果の平均は以下の 5 のようになった．また，1 件あたりの処理時間は平均 41μs であった．

5. 考察

実験結果から，危険度の閾値を増加するに伴い，TP Rate が現象し，TN Rate が増加していくことがわかる．フィッシングサイトと正規サイトの危険度の分布は図 1 のようになる．図 1 から分かるよう，フィッシングサイトと正規サイトの危険度の分布は大きく被っているため，判別の精度を向上することができなかったと考えられる．次に各パラメータについて考察していく．

正規サイトとフィッシングサイトの階層数の分布の割合は，図 2 のようになる．98%を超える正規サイトの階層数が 1 である．一方階層数が 1 のフィッシングサイトは，5 割程度である．そのため本手法で用いた，TLD または ccTLD とそれに紐づくセカンドレベルドメインを除いた

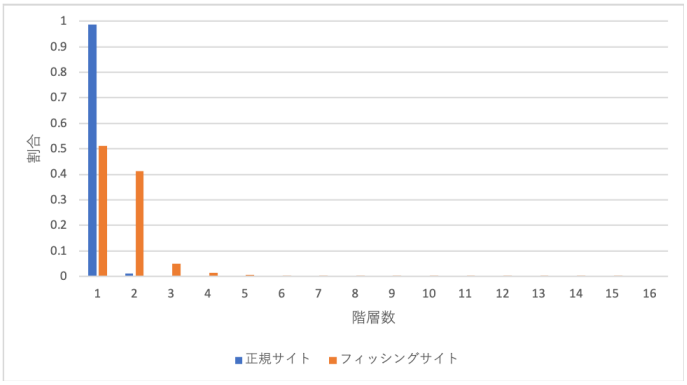

図 2　正規サイトとフィッシングサイトの階層数の分布
Fig. 2　Distribution of hierarchy

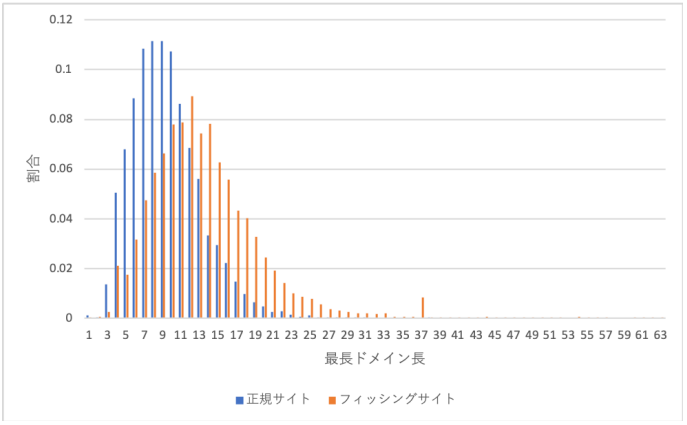

図 3　正規サイトとフィッシングサイトの最長ドメイン長の分布
Fig. 3　Distribution of the longest domain length

階層数は，フィッシングサイトの判別の要素の 1 つとして有効であると考えられる．

正規サイトとフィッシングサイトの最長ドメイン長の分布の割合は，図 3 のようになる．正規サイトよりもフィッシングサイトの方が最長のドメイン長が長いことが分かる．しかし，正規サイトとフィッシングサイトの分布の多くが被っているため，この要素を利用すると誤検知が増加すると考えられる．

正規サイトとフィッシングサイトの数字の塊の数の分布の割合は，図 4 のようになる．正規サイトの 99%以上が数字の塊が 1 個以下である．そのため，数字の塊の個数を利用してフィッシングサイトの検知を行う場合，誤検知が起こりにくいと考えられる．よって，数字の塊の数はフィッシングサイトの検知に有効な要素と考えられる．

正規サイトとフィッシングサイトのハイフンの数の分布の割合は，図 5 のようになる．正規サイトの 99%以上がハイフンの数が 1 個以下である．そのため，ハイフンの個数を利用してフィッシングサイトの検知を行う場合，誤検知が起こりにくいと考えられる．よってハイフンの数をフィッシングサイトの検知の要素として有効であると考えられる．

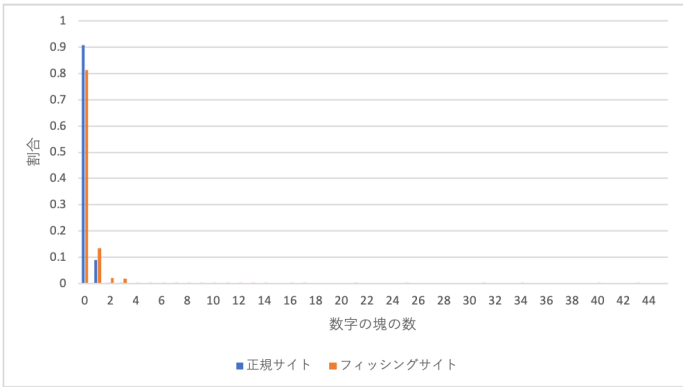

図 4　正規サイトとフィッシングサイトの数字の塊の数の分布
Fig. 4　Distribution of number of lump of number

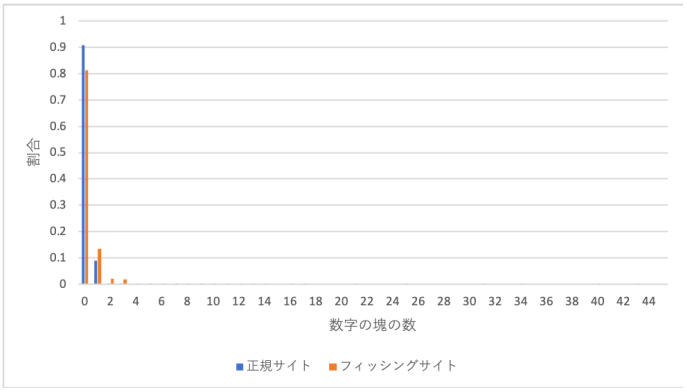

図 5　正規サイトとフィッシングサイトのハイフンの数の分布
Fig. 5　Distribution of number of hyphen

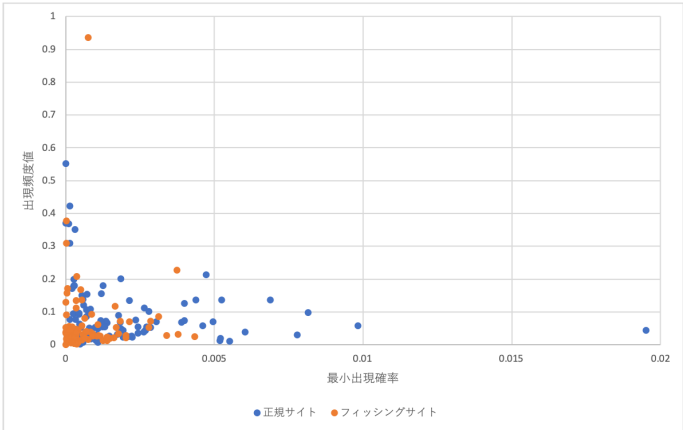

図 6　出現頻度値-最小出現確率の散布図
Fig. 6　Scatter plot: Appearance frequency value - minimam probability of appearance

縦軸に出現頻度値，横軸に最小出現確率を取った，正規サイトとフィッシングサイトの散布図を図 6 示す．最小出現確率に関しては，正規サイトの方がフィッシングサイトよりも分布が大きい方に寄っている．しかし，値の小さい部分に正規サイトとフィッシングサイトが混在しているため，検知に用いることで誤検知が増加すると考えられる．出現頻度値に関しては，正規サイトとフィッシングサイトの間に違いがほとんどなかった．今回，出現頻度を文字ご

とではなく，文字を 5 つのグループに分割して出現頻度の計測を行った．しかし，URL の文字列は非常に短いため，5 つのグループでもグループ数が多かった可能性がある．さらにグループ数を減らして実験を行い，要素の妥当性を調べる必要があると考えられる．

6. まとめ

本論文では，迅速で誤検知の少ないフィッシングサイトの検知を目的とし，URL 文字列に着目して検知を行った．そこで，有効となり得る要素について検討を行った．その結果，ドメインの階層数，数字の塊の数，ハイフンの数が有効であると判断した．

今後の課題として，今回利用した出現頻度値と最小頻度確率について再検討を行い，改善する必要がある．また，複数要素の組み合わせによって危険度を決定するような仕組みも検討することでさらなる精度の向上を目指す．今回は危険度を加算する一方であったが，信頼できる要素がある場合に危険度を減らすような仕組みを考える，といったことが挙げられる．

参考文献

[1] Jun Ho Huh and Hyoungshick Kim: *Phishing Detection with Popular Search Engines: Simple and Effective.*, Springer-Verlag, LNCS 6888, 194-207, 2012.
[2] Zhang Yue, Hong Jason and Cranor Lorrie: *CANTINA: A content-based approach to detecting phishing web sites*, Proceedings of the 16th international conference on World Wide Web, 639-648, 2007.
[3] Anti-Phishing Working Group: *Phishing Activity Trends Report for the Month of February, 2007*, 入手先 ⟨https://docs.apwg.org/reports/apwg_report_february_2007.pdf⟩, （参照 2019-08-17）.
[4] *2018 Phishing and Fraud Report: Attacks Peak During the Holidays* 入手先 ⟨https://www.f5.com/labs/articles/threat-intelligence/2018-phishing-and-fraud-report–attacks-peak-during-the-holidays⟩, （参照 2019-08-17）.
[5] Shahid Shah: *Statvoo Top 1 Million Sites*, Opsfolio Community, 入手先 ⟨https://www.opsfolio.com/resourcecenter/statvoo-top-1-million-sites/⟩, （参照 2019-06-12）.
[6] *PhishTank — Join ths fight against phishing*, 入手先 ⟨https://www.phishtank.com/⟩, （参照 2019-08-17）.
[7] *mozilla wiki TLD List*, 入手先 ⟨https://wiki.mozilla.org/TLD_List⟩, （参照 2019-08-05）.