

模倣コンテンツの特性に基づくフィッシング検知方式

中山 心太† 吉浦 裕‡

†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

‡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

E-Mail: †shinta@edu.hc.uec.ac.jp ‡yoshiura@hc.uec.ac.jp

概要 偽のウェブサイトを用いて個人情報などを不正に入手するフィッシング詐欺が増加している。本論文では、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、フィッシングサイトと正規サイトの類似性に基づくフィッシング検知方式を提案する。提案方式は、ホワイトリストおよびブラックリストを必要としないのが特徴である。実際のフィッシングサイト及び正規サイトを用いて、提案方式の検知能力を評価する。

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

Phishing detection based on features of mimic content.

Shinta Nakayama† Hiroshi yoshiura‡

†The Department of Human Communication, The Graduate School of Electro-Communications,
The University of Electro-Communications

‡The Department of Human Communication, The Faculty of Electro-Communications,
The University of Electro-Communications

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

E-Mail: †shinta@edu.hc.uec.ac.jp ‡yoshiura@hc.uec.ac.jp

Abstract Phishing, the fraud to get personal information illegally through fake Web sites, is prevailing. This paper proposes a new phishing detection method based on the similarity between a phishing site and its original site because the phishing site is a mimic of the original. The proposed method does neither require a while list nor a black list. The method has been evaluated by using actual phishing sites and their originals.

Keywords Phishing attack, Network security, web

1. はじめに

インターネットの普及に伴い、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。2006年度のインターネット世帯普及率は85.4%[1]であり、低リテラシー層の利用が広がっていることが伺える。それに伴い、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作成し、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は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、フィッシングサイトの運営されている

サーバの特性を利用した方式などがある。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ったり、個人が運営しているサーバなどを誤検知してしまう可能性がある。

そこで本研究では、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、コンテンツレベルの類似性に基づいてフィッシング詐欺検知を行う技術を提案する。

2. 従来のフィッシング対策研究

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、載っていないサイトを弾く方式である[2]。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。

2.2 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[3]。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。近年ではマルウェアによってBot化されたコンピュータ上でフィッシングサイトが公開されることが多く、同じ内容のフィッシングサイトが一瞬のうちに、多数のURLで公開される。そのため、ブラックリストへの登録漏れが発生する可能性が高まっている[4]。

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の特性を利用したものがある[5]。米国のAPWG(Anti Phishing Working Group)の調査によると、フィッシングサイトの平均存続期間は4日と非常に短い[6]。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。しかし、個人サーバや、新たにできたウェブサイトをフィッシングサイト扱いしてしまう可能性がある。

2.4 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を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[7]。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、ユーザーサイドで判定することはできない。

2.5 ユーザーの認知能力の分析

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[8]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、もっと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、ユーザー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わかる。

3.模倣コンテンツの特性に基づくフィッシング検知方式

3.1 基本アイデア

フィッシングサイトの多くは正規サイトを模倣して作られたものである。したがってフィッシングサイトと正規サイトには、何らかの類似性が存在するはずである。文章やデザインがそのままのコピーであるとは限らないが、特徴的な表現であったり、表している企業名であったり、何らかの広い意

味での類似性が存在すると考えられる。

したがって、フィッシングサイトは以下の本質的特徴がある。

- 模倣された正規サイトが存在している。
- フィッシングサイトと正規サイトは、そのままコピー、特徴的な表現、企業名など、多岐にわたるが、広い意味では類似している。

そのため、現在閲覧している疑わしいサイトの特徴情報を抽出し、それ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できる。そして、現在閲覧している疑わしいサイトと、発見した正規サイト候補との整合性を調べることでフィッシングサイトかどうかを判定することができる。以上のアイデアをコンピュータが実装するには次の3ステップが必要である。図1はこれを図式化したものである。個々の内容の詳細は次節で説明する。

① 認知処理による類似情報の抽出

現在閲覧している疑わしいサイトの内容を読み取って、正規サイトと類似しているであろう情報(以下、類似情報と呼ぶ)を認識する。

② 正規サイト候補の検索

類似情報をキーワードとして、検索エンジンで正規サイト候補を検索する。

③ 類似性の比較

正規サイト候補とフィッシングサイト候補との内容の類似性と、URLを比較して、フィッシングサイトかどうか判断を行う。

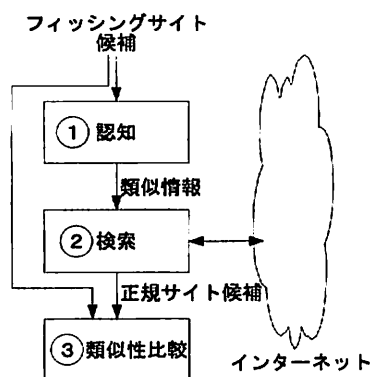


図1 基本アイデアの概要

3.2 認知による類似情報の抽出

3.2.1 類似性の分析

本研究では、コンピュータがフィッシングサイト候補を認知して、内容を解析し、類似情報を抽出しなければならない。一口に類似といっても、データレベル、要素レベル、構造レベル、意味内容レベルなど、様々なレイヤーの類似性が存在する。そのため、要素、構造、意味といった、様々な抽象レベルで類似情報をデータから取り出さなくてはならない。それぞれの階層と実装容易性、多様性の吸収性を図2に示す。

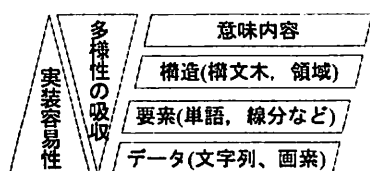


図2 類似性の階層

一般に上位層になるほど、表面的に異なっているが、意味内容の類似性を正確に捉えることができる。たとえば「電気通信大学」と「調布の理工系大学」は表面上はまったく異なる単語であるが、意味はほとんど同じである。

しかし、上位層の解析ほど技術が確立しておらず、また汎用的なツールも整備されていないため、実装が難しい。また、情報の意味を理解するには、金融や流通などそれぞれの分野に依存した知識が必要であり、汎用性に乏しく、メンテナンスのコストがかかる。そのため、上位層の認知を直接行わず別の手段で同等の効果を得ることが課題である。

3.2.2 統計分析の利用

自然言語処理の分野では、統計的手法により、単語の意味を分析する手法が研究されている。たとえばTF-IDF法と共起分析[9]がある。TF-IDF法とは、ある文章中の単語の出現頻度と珍しさidfとの積によって、単語の重み付けを行い、その文章で特徴的な単語を調べるアルゴリズムである。また共起分析はサンプルテキスト中に二つの語が同時に現れる頻度を測定することで、語の近さを求めるアルゴリズムである。

そこで本研究では、統計分析を利用して意味理解を補うことを考える。統計分析をするには膨大なサンプルテキストが必要になる。またサンプルテキストはその時代を反映したものであることが望ましい。そのため、これを満たすために、検索によってウェブ上のテキストをサンプルテキストとして用いる。

3.3 正規サイト候補の検索

検索エンジンは検索結果が表示される際に、なんらかのアルゴリズムで、順位付けを行って出力している。たとえばGoogleではPageRankと呼ばれるアルゴリズムが利用されている[10]。このアルゴリズムは大小様々なルールで規定されているが、もっとも比重が大きいルールに次の2つがある。

- 多くのページからリンクされているページは良質である。

- 良質なページからリンクされているページは良質である。

フィッシングサイトは基本的にスパムメールに記載されて、ユーザーに広まるため、ウェブページからリンクが張られることは少ない。また、仮にフィッシングサイトにリンクを張るようなサイトがあったとしても、そのようなサイトが優良であるはずがない。そのため、フィッシングサイトの評価は非常に低くなると考えられる。

一方、正規サイトは提携会社や多くの利用者からリンクされているため、評価は高い。そのため企業名を検索した際に、正規サイトよりもフィッシングサイトが検索結果の上位に来ることは考えにくい。そのため、検索エンジンを適切に用いることで正規サイト候補を見つけ出せる。

3.4 類似性の判定

得られた正規サイト候補について、フィッシングサイト候補との比較を行う。比較はURLの比較と、内容の比較を行う。表1は比較方法を表にしたものである。

表1 URLと内容の比較によるフィッシング判定

	内容が一致	内容が不一致
① URL が一致	正規サイト	アドレスバーが書き換えられている
② URL が不一致	フィッシングサイト	統計分析の失敗

① URL が一致しているとき

正規サイトのサーバが乗っ取られない限り、フィッシングサイトと正規サイトとが同じドメインで運営されていることは無い。そのため、正規サイト候補と、フィッシングサイト候補のURLが一致していたら、正規サイトであるとみなすことができる。しかし内容が一致していなければ、ブラウザの脆弱性を突かれて、アドレスバーが書き換えられている可能性がある。

② URL が一致していないとき

URLが不一致で、内容が一致していた場合、フィッシングサイトだとみなすことができる。しかしフィッシングサイトは利用者に個人情報を入力を促す文章などが追記されていることが多いため、内容の比較はある程度の揺らぎを吸収しなくてはならない。2つの文章間の類似度判定はTF-IDF法とベクトルの内積を利用した手法などが提案されている。

URLも内容も一致していなかった場合、統計分析が失敗し、正規サイトと類似しているキーワードの抽出に失敗したため、正規サイトが正しく検索できなかったことがわかる。

そのため、統計分析の結果を再評価して次善のキーワードを元に検索を行い、フィッシング判定のやり直しを行う。さらに利用者に正しい意味を問い合わせることで、判定の精度を向上させることができる。

4.実装

3章の方針に従い、フィッシングサイト候補のテキストを形態素解析[11]し、統計分析を行うことによって、構文解析、意味解析の代わりとする。これによって、要素レベルの情報から、意味内容レベルの情報を持ったキーワードをN個を抽出する。このN個のキーワードを用いて検索を行い、正規サイトの候補をM件得る。M件の正規サイト候補に対してそれぞれ類似判定を行い、フィッシングサイトかどうか判定する。図3は以上を図式化したものである。

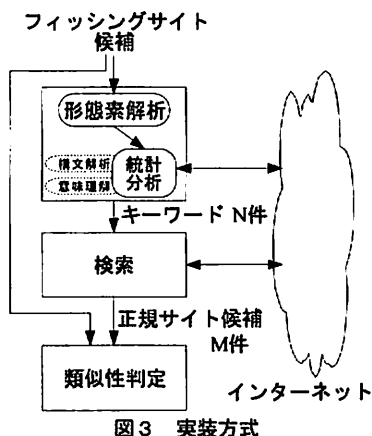


図3 実装方式

4.1 処理の流れ

形態素解析エンジンにはmecab[12]、辞書にはipadic[13]を採用した。統計分析にはTF-IDF法を用いた。検索エンジンにはGoogleを用いた。

試作システムは次のような流れになっている。

- 1 フィッシングサイト候補からテキストを取り出す。
- 2 形態素解析を行い、名詞を抽出し、出現回数tfを計測する。
- 3 得られた名詞をキーワードとして検索し、ヒット件数からidf値を計算し、名詞のTF-IDF値を求める。
- 4 TF-IDF値の上位N個の名詞をキーワードとして検索を行う。
- 5 M件の検索された正規サイト候補と本システムに入力されたフィッシングサイト候補とを比較し、フィッシング判定を行う。

4.2 TF-IDF法の実装

TF-IDF法は文章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手法であり、次の式で表される。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。

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

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

TF-IDF法は検索エンジンを用いて実装する手法が提案されている[14]。検索エンジンが公表している、インデックスしているページの総数をS、検索エンジンでtを検索したときの検索数をdf(t)とすることで、idf(t)を求めることができる。これによりローカルにidfのデータベースを持たなくとも、単語の重み付けを行うことができ、また新語や、ネットワーク上の情報の変化に対応することができる。

Yahoo!の広報資料[15]からS=192億として実装を行った。

表2は三菱東京UFJ銀行のトップページをTF-IDF法を用いて計算した、TF-IDF値の上位10個の名詞である。高い精度でサイトの意味を捕らえていることがわかる。

ただしidfの計算のために名詞を検索した際に1件も検索できなかった場合、idfが無限大に発散してしまう。今回の研究はフィッシングサイトの元となるサイトを検索することが目的であるため、検索できないキーワードは必要としない。そのため、そのようなキーワードのidfは0とした。

表2 TF-IDFの算出例

名詞	TF-IDF値
一覧	59.14
三菱	50.44
保険	49.00
口座	44.66
UFJ	43.34
当行	42.86
証券	39.03
お知らせ	39.03
入会	38.30
国債	34.29
抜粋	31.88

4.3 類似判定の方法

今回の実装では、内容の比較は行わず、URLの比較のみを行った。URLの一致度には次の種類がある。

- ① 完全一致
- ② ディレクトリまで一致
- ③ サブドメインまで一致
- ④ ドメインまで一致
- ⑤ ドメインが不一致

正規サイトのサーバが乗っ取られない限り、フィッシングサイトのドメインは正規サイトとは別のものである。そのため、正規サイト候補M件のうち1件でも『④ドメインまで一致』以上があれば、正規サイトとみなすようにした。

5.評価

5.1 フィッシングサイトの評価と考察

サイトのテキストから抽出するキーワードの件数Nを10、検索される正規サイト候補の件数Mを10とした。評価期間中に捕捉することができたフィッシングサイト4件について、本システムを用いて実験を行った。評価した結果4件中3件はフィッシングサイトであると検知した、1件は正規サイトであると誤検知した。

Case1: 某銀行インターネット事業部

TF-IDF のキーワードの上位 10 件は次のようになった。

「時間 場合 日曜日 営業 銀行 時 利用 分 店舗 ATM」これらのキーワードで検索したところ、上位 10 件は次のようになった。

- 1 http://www.XXXXXX.co.jp/fee/atm_cd.html
- 2 http://www.XXXXXX.co.jp/loan/card/
- 3 http://www.XXXXXX.co.jp/direct/time.html
- 4 http://www.XXXXXX.co.jp/fee/time_conveni_atm.html
- 5 http://www.XXXXXX.co.jp/loan/card/index_2.html
- 6 http://www.XXXXXX.co.jp/useful/atm/conveni_atm.html
- 7 http://www.ZZZZZZ.co.jp/kojin/time/index.html
- 8 http://www.ZZZZZZ.co.jp/kojin/direct/jikan/index.html
- 9 http://www.ZZZZZZ.co.jp/kojin/tenpo/atm/honshiten/index.html
- 10 http://www.YYYYYY.co.jp/salut/osirase.html

www.XXXXXX.co.jp は正規サイトのドメインである。以上から、銀行名などこのサイトの意味を表すキーワードは正しく捕らえていないが、銀行の正規サイトを捉えることができたことがわかる。また、フィッシングサイトのドメインを含むページが検索されなかったため、4.3 節の評価ルールから、フィッシングサイトであると判定した。

検索された正規サイトの中身を調べると、ATM の利用案内のページであった。フィッシングサイトにもまったく同様の記述が存在した。そのため、フィッシングサイトがこの利用案内をそのままコピーして利用していたため、正規サイトを検索できたものと考えられる。

Case2: 某クレジット会社

TF-IDF のキーワードの上位 10 件は次のようになった。

「申込 審査 収入 歳 会員 <サービス名> 以降 了承 ローン 借入れ」(<サービス名> は正規のサービス名)

これらのキーワードで検索したところ、上位 10 件に正規サイトの URL を得ることができなかった。そのため、4.3 節の評価ルールからフィッシングサイトだと判定した。

しかし、検索キーワード 10 件中正規サイトに存在しないキーワードが含まれていたため、正規サイトが検索できなかった。正規サイトのサービス名という正しいキーワードは取得できているため、キーワード選定の高度化を行うことで、正規サイトを検索できる可能性がある。

Case3: 某オークションサイトの ID 更新手続きページ

TF-IDF のキーワード上位 10 件は次のようになった。「英数字 入力 半角 登録 番号 数字 <企業名 1> <企業名 2> セキュリティー 手続き」(スペースの入った企業名であったため、企業名が <企業名 1> <企業名 2> の二つに分割された)

これらのキーワードで検索したところ、1,2,8 位に親会社の URL が、3~7,10 位に正規サイトが検索された。そのため、4.3 節の評価ルールからフィッシングサイトだと判断した。

Case4: 企業グループの子会社のクレジットカード会社

TF-IDF のキーワードの上位 10 件は次のようになった。

「<企業グループ名> 名前 記入 年収 融資 診断 審査 漢字 携帯 登録」(<企業グループ名> は正規の企業グループ名)

これらのキーワードで検索を行ったところ、フィッシングサイトの URL が検索結果の 1 位に来てしまった。

4.3 節の評価ルールから、正規サイトだと判定されてしまった。これはフィッシングサイトが長期間生存していて、検索エンジンに登録されてしまっていたこと、サイトが一から作られていたため、企業グループ名以外のキーワードが正規サイトとほとんど一致しなかったためである。しかしフィッシングサイトのタイトルには正規企業名が書かれていた。TF-IDF 法とともに、HTML のタグによるキーワードの重み付けを行い、title や h1 タグなどに含まれる名詞に重み付けすることによって、グループ会社名だけでなく、企業名も抽出できる可能性があった。

5.3 正規サイトの評価と考察

正規サイトが正規サイトであると検出されるかどうかの実験を行った。実験に利用したのは国内銀行 22 社のトップページと、国内銀行 17 社ネットバンクのログインページの計 39 件である。その結果が表 3 である。

表 3 正規サイトの検出実験結果

URL 一致度	件数
①完全一致	26
②ディレクトリまで一致	1
③サブドメインまで一致	1
④ドメインまで一致	2
⑤ドメインが不一致	9

正規サイトの検出には 39 件中 9 件が失敗した。その理由には次のようなものが挙げられる。

- 形態素解析の失敗などにより、検索エンジンにインデックスとして選定されていない単語をキーワードとしてしまったため。
- ネットバンクのログインページなどの認証ページは、一般のトップページから移動することを前提にしており、検索エンジンに登録しないように META タグが記述されていたため。
- トップページが更新直後で、更新されたニュースの中の項目をキーワードとして捉えてしまい、検索エンジンのキャッシュ中にそのキーワードが存在しなかったため。
- ログインページが銀行自身のドメインとは別のドメイン (anser.or.jp NTT データが管理する地方銀行用のネットバンクサービス) で運営されており、銀行自身の URL を検索してしまったため。

TF-IDF 値の上位 10 個のキーワードの中に、形態素解析の失敗によって生まれたキーワードや、検索エンジンにインデックスされていないキーワードが含まれているため、検索に失敗したことがわかる。そのようなキーワードを除去するために、キーワード 10 個中から 8 個の組み合わせで検索を行うようにしたところ、失敗した 9 件中 8 件で正規サイトと同じドメインのページが検出できた。そのため、正規サイトと判定することができた。また、残る 1 件についても、N=11 として TF-IDF の上位 11 個で検索を行うと、正規サイトを検索することができた。よって、形態素解析の精度と、キーワード選定の精度が向上すれば、誤検出を防止できることがわかった。

6.まとめ

本研究では、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、フィッシングサイトと正規サイトの類似性に基づくフィッシング検知方法を提案した。本手法ではフィッシングサイト候補から、自然言語処理と統計分析によってキーワードを抽出する。そして抽出されたキーワードを用いて、検索エンジンで正規サイトを見つけ出し、正規サイトとの比較によりフィッシングサイトであると判定する。これによりデータベースに依存しない

実際にフィッシングサイトにアクセスして評価実験を行ったところ、4 件中 3 件はフィッシングサイトであると判断し、1 件は正規サイトであると誤検知した。

正規サイトにアクセスして実験を行ったところ、39 件中 30 件が正規サイトと判断した。失敗した 9 件の失敗原因を分析したところ、キーワードの選び方が失敗しているケースがほとんどであったため、キーワード選定方法を改良したところ、すべてのケースで正規サイトであると判定することができた。

今後の課題として、以下の内容を検討する。

- ・タグ情報の反映、共起分析によるノイズキーワードの除去など、TF-IDF 以外の手法を検討することで、キーワードの選定方法を高度化する。
- ・現在は TF-IDF の計算に名詞 1 件当たり 0.3 秒程度かかっている。ウェブページには 300 前後の名詞が含まれているため、ひとつのページを解析するのに 3 分程度かかっている。そこで TF-IDF の結果をキャッシュして高速化を図る。
- ・今回は日本語のサイト用のシステムを開発したため、英語のサイトには対応していない。そのため、十分なフィッシングサイトのサンプルを得ることができなかった。今後英語サイトへの対応をし、十分な数のサンプルで本提案が有効に働くか実験する。
- ・ブラックリスト、ホワイトリストなどの他手法との併用を検討する。

謝辞

本研究は電気通信大学 SVBL2006 年度 学生・一般アイデアコンテスト及び、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。

参考文献

- [1] 財団法人インターネット協会監修,インターネット白書 2006,株式会社インプレス R&D,pp.38-41
- [2] 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
- [3] "RBL.JP プロジェクト",<http://www.rbl.jp/>
- [4] "フィッシング詐欺サイト情報 リファレンス I [ゾンビ PC 詐欺] 53.com など複数同時詐欺",
<http://www.rbl.jp/phishing/index.php?UID=1163579393>
- [5] 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" IPSJ DSP-126 CSEC-32 2006 年 3 月,pp.13-18
- [6] Anti-Phishing Working Group"APWG Phishing Trends Activity Report for February 2007.",
http://www.antiphishing.org/reports/apwg_report_february_2007.pdf
- [7] Liu Wenyin ,Guanglin Huang ,Liu Xiaoyue ,Zhang Min ,Xiaotie Deng "Detection of Phishing Webpages based on Visual Similarity",Proc. of WWW2005 ,pp1060 - 1061
- [8] Rachna Dhamija,J. D. Tygar,Marti Hearst ,"Why Phishing Works." CHI2006
- [9] Julie Weeds, David Weir, "Co-occurrence Retrieval: A Flexible Framework for Lexical Distributional Similarity" Computational Linguistics Vol. 31, No. 4, pp.439-475
- [10]Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", 1998,<http://dbpubs.stanford.edu:8090/pub/1999-66>
- [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] "McCab Yet Another Part-of-Speech and Morphological Analyzer",<http://mccab.sourceforge.net/>
- [13] "ipadic",<http://sourceforge.jp/projects/ipadic/>
- [14] "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>
- [15] "Yahoo! Search Blog: Our Blog is Growing Up・And So Has Our Index",<http://www.ysearchblog.com/archives/000172.html>