2-4　スパムによる攻撃を分析するためのクラスタリング手法と特徴量選択手法について

2-4 Clustering and Feature Selection Methods for Analyzing Spam Based Attacks

宋　中錫
SONG Jungsuk

要旨

近年になってスパム・メールが激増し、スパムはインターネット上の深刻な脅威として認識されている。最近のスパム・メールは、ボットが他のボットと結合してボットネットを構成することで送信され、スパーマーはスパムの分析機能や検出テクノロジーによって検出されることを巧みに回避しようとしている。さらに、ほとんどのスパム・メッセージにはスパムの受信者を悪意のあるウェブ・サーバーに誘導するURLが含まれており、マルウェアの感染やフィッシング攻撃のような様々なサイバー攻撃を実施しようとしている。スパムによる攻撃に対処するために、電子メールの類似度に基づいてスパム・メールのクラスタリングを行おうとする取り組みが数多く存在する。スパム・メールのクラスタリングによって得られるスパム・クラスタを使用して、スパムを送信するメカニズムと悪意のあるウェブ・サーバーのインフラの見極めを行い、それらの要素がどのように連携しているのかを検出することができる。上記に基づいて、O-means法という最適化されたスパム・クラスタリング手法を提案する。これは、最も広範に使用されているクラスタリング手法であるK-means法に基づくものである。SMTPサーバーで収集した3週間分のスパムを分析することによって、O-means法によるクラスタリングは87%の確度を実現することを確認した。これは、従来のクラスタリング手法よりも優れている。また、スパム・メール間の類似度の比較を行うために12種類の統計に基づく特徴量を新たに定義することによって、O-means法によるクラスタリング手法の効果を増大するために役立つ、最適化された一連の特徴量を検出する特徴量選択の手法を提案する。本手法を活用することで、計算時間を短縮したうえで86.33%のクラスタリング確度を実現する4種類の重要な特徴量の特定を行うことができた。

In recent years, the number of spam emails has been dramatically increasing and spam is recognized as a serious internet threat. Most recent spam emails are being sent by bots which often operate with others in the form of a botnet, and skillful spammers try to conceal their activities from spam analyzers and spam detection technology. In addition, most spam messages contain URLs that lure spam receivers to malicious Web servers for the purpose of carrying out various cyber attacks such as malware infection, phishing attacks, etc. In order to cope with spam based attacks, there have been many efforts made towards the clustering of spam emails based on similarities between them. The spam clusters obtained from the clustering of spam emails can be used to identify the infrastructure of spam sending systems and malicious Web servers, and how they are grouped and correlate with each other. Therefore, we propose an optimized spam clustering method, called O-means, based on the K-means clustering method, which is one of the most widely used clustering methods. By examining three weeks of spam gathered in our SMTP server, we observed that the accuracy of the O-means clustering method is about 87% which is superior to the previous clustering methods. In addition, we define new 12 statistical features to compare similarities between spam emails, and we propose a feature selection method to identify a set of optimized features which makes the O-means clustering method more effective. With our method, we identified 4 significant features which yielded a clustering accuracy of 86.33% with low time complexity.

[キーワード]
スパム，クラスタリング，特徴量選択，ボットネット，悪意のあるURL
Spam, Clustering, Feature selection, Botnet, Malicious URLs

1 はじめに

近年になってスパム・メールが激増し、スパムはインターネット上の深刻な脅威として認識されている。最近のスパム・メールは、ボットが他のボットと結合してボットネットを構成することで送付され、スパーマーはスパムの分析機能や検出テクノロジーによって検出されることを巧みに回避しようとしている。例えば、ボットネットの所有者はボットを慎重に制御することで、各ボットから微量のスパムを送付することによって、リモートシステムが送付する一定量のトラフィックのチェックを行う従来のボット検出テクノロジーを回避しようとしている。最近のボットは平均で1件から2件のスパム・メッセージを送付するために使用されており[1]、作者による実験では各スパム送信システムは平均で1.9件の電子メールを送信していることを確認した。さらに、ボットが効力を発揮する時間と特定のスパム・メッセージの有効期間は非常に短い。一部の推定値によると、ボットの75%が効力を発揮する時間は2分以下で、スパム・メッセージの65%が効力を発揮する時間は2時間以下である[2][3]。

一方、ほとんどのスパム・メッセージにはスパムの受信者に悪意のあるウェブ・サーバーにアクセスさせることで、マルウェアの感染やフィッシング攻撃等の様々なサイバー攻撃をしかけようとするURLも含まれている。したがって、あるウェブ・ページが悪意のあるものかどうかを確認するために、URLに関連付けられたウェブ・ページを分析する必要がある。しかしながら、ほとんどの場合全てのウェブ・ページをリアルタイムで分析することは非常に困難である。現在送信される電子メールの90%以上はスパムと考えられ[4]、様々な目的で使用されているためである。さらに、実際に使用される悪意のあるウェブ・サーバーはURLに関連付けられた最初のウェブ・ページから4つ以上のウェブ・ページに影響を及ぼしている[5][6]。したがって、スパムに基づく攻撃を分析する所要時間を短縮する必要がある。

スパムに基づく攻撃に対応するために、スパム・メール間の類似度によってスパム・メールのクラスタリングを行おうとする取り組みが数多く存在する[1][7]-[10]。スパム・メールのクラスタリングを行うことで、共通する類似度に基づいてスパムをクラスタに分類することができる。各スパム・クラスタを分析することで、スパム・メールを送信するために使用されるシステムのインフラを見極め、スパム・メールがどのように分類されているのかを把握し、スパムを送信するシステムと悪意のあるウェブ・サーバーの間の関連性を特定することができる。スパムに基づく攻撃においては攻撃者が効果的に攻撃を行うためにはこれらの準備を行う必要があるためである。さらにスパムの分類において得られるクラスタ情報を使用することで、ウェブ・ページの分析に必要な時間を最小化することができる。つまり、ある電子メールが特定のクラスタに属する場合、URLをたどってウェブ・ページの分析を行う必要がなくなるということを意味する。したがって、スパム・メールのクラスタリングが効果的にスパムに基づく攻撃を分析するために不可欠であることが言える。また、より正確にスパムに基づく攻撃を分析するためには、スパムのクラスタリングの確度を改善することが重要になる。

O-means法というスパムの最適化されたクラスタリング手法を本論文で提案する。これは最も広範に使用されているクラスタリング手法の1つであるK-means法[11]に基づくものである。K-means法によるクラスタリング手法は、データセットを以下のステップに基づいてK個のクラスタに分類する。

- 初期設定：あるデータセットからK個のインスタンスをランダムに選択し、初期のクラスタの中心とする。
- 割り当て：各インスタンスを最も近い中心に割り当てる。
- 更新：各クラスタの中心を当該クラスタのメン

バーの平均値で置き換える。
- 繰り返し：各クラスタにおける値が変化しなくなるか、終了条件を満たすまで、割り当てと更新を繰り返す。

K-means法によるクラスタリング手法に人気があるのは、計算時間が短縮でき、簡単で、迅速に値の収束を行えるためと言える。しかし、このクラスタリング手法による結果は選択されたK個の初期の中心によるところが大きく、適切な数（K個）のクラスタを事前に設定することが非常に難しいことも分かっている。O-means法によるクラスタリング手法では、K-means法によるクラスタリング手法の欠点を克服することでパフォーマンスを向上することができる。SMTPサーバーで収集した3週間分のスパムを分析することによって、O-means法によるクラスタリングでは87%の確度を実現することを確認した。これは、従来のクラスタリング手法よりも優れている。また、スパム・メール間の類似度の比較を行うために12種類の統計に基づく特徴量を新たに定義することによって、O-means法によるクラスタリング手法の効果を増大するために役立つ、最適化された一連の特徴量を検出する特徴量選択の手法を提案する。本手法を活用することで、計算時間を短縮したうえで86.33%のクラスタリング確度を実現する4種類の重要な特徴量の特定を行うことができた。

本論文は以下のとおり構成する。**2**では従来のスパムのクラスタリング手法について簡単に説明する。**3**では今回提案するクラスタリング手法について説明し、**4**と**5**ではそれぞれ実験の結果と考慮ポイントについて記載する。最後に、**6**では結論と将来への提言について述べる。

2　関連研究

Zhuangをはじめとする研究者は、スパム・メールのデータを使用することでボットネットのメンバー情報のマッピングを行う新規の手法を構築した[8]。このクラスタリング手法は、同様のコンテンツを含むスパム・メールは同一のスパマーまたは攻撃者によって送信されていることが多いという前提に基づいている。電子メールのメッセージが共通した経済的な目的に基づいていることに着目したためである。その結果、同様のスパム・メッセージや関連するスパム・キャンペーンを分類することで数百のボットネットを特定することに成功した。Liをはじめとする研究者は、自ら構築したドメイン・メール・サーバーで収集したスパムのトラフィックに基づいてスパマーのクラスタリング構造の研究を行った[7]。この手法においては、スパム・メールに含まれるURLをクラスタリングの基準として使用し、スパマー間のリレーションシップを研究することで高度なクラスタリング構造が判明したことを発表した。文献[1]においては、Xieをはじめとする研究者はスパムを発信するボットネットの特徴を分析した。まず、同じドメイン名を持つ様々なURLを使用してスパム・メールの分類を行った。その結果、HotmailやAutoREをはじめとする独自のシステムに含まれる3か月分の電子メール・サンプルに基づいて7,721件のボットネットによるスパム・キャンペーンと340,050個のボットネット・ホストのIPアドレスを特定した。しかし、スパム・メッセージやトレンドの変化によってコンテンツ、URL、およびドメイン名の3つの基準が簡単に影響を受けるため、この手法には致命的な欠陥がある。実際、スパマーは定期的に電子メールのコンテンツとドメイン名の変更を行い[12]、URLの有効期間は非常に短くなっている（1日間：45%、2日間：20%、3日間：7%）[1]。また、作者による実験では、スパム・メールで使用されるURLのほとんどはユニークなものであることを確認している。その結果、これら3つの基準はクラスタの品質を維持または向上するのに適していないということが言える。

文献[9]および[10]において、URLから抽出したIPアドレスに基づいてスパム・メールのクラスタリングを検証する実験を行った。つまり、URLから抽出した2つの電子メールのIPアドレスが全く同一の場合、当該の2つの電子メールを同じクラスタとして認識した[9]。本手法によるパフォーマンスはドメイン名とURLに基づくクラスタリング手法よりも優れているものの、完璧とは言えない。IPクラスタには、様々な制御を行うエンティティから送信され、含まれるURLが様々な種類のウェブ・ページに接続する相互に関連性のない大量の電子メールが含まれているためである。その理由としては、多くのウェブ・サーバーが同じIPア

ドレス上で多くのウェブサイトのホスティングを行っており、スパーマーや攻撃者によって攻撃される多くのウェブ・サーバーが独自のウェブサイトを構築していることが挙げられる。本論文では、同じIPクラスタに属するウェブサイトを個別のクラスタに分類することについて主に説明する。

3 提案手法

3.1 プロシージャの概要

図1は、本論文にて提案するO-means法によるクラスタリング手法のプロシージャの概要を示している。本プロシージャは以下の6つの主な要素で構成される。

- ヘッダーに基づくクラスタリング機能：**3.2**で説明する電子メールのヘッダー間の類似度に基づいて、特定のスパム・メールからHDクラスタ（例：HD_C_1、HD_C_2、…HD_C_h）を構築し（①）、O-means法のクラスタリング機能に挿入する（②）。
- IPに基づくクラスタリング機能：スパム・メール内のURLからIPアドレスを抽出し、**3.3**で説明するように抽出したIPアドレスを使用してスパム・メールからIPクラスタ（例：IP_C_1、IP_C_2、…IP_C_i）を生成し（③）、O-means法のクラスタリング機能に挿入する（④）。
- O-means法によるクラスタリング機能：**3.4**で説明するように、K-means法によるクラスタリング手法に基づいてスパム・メールからOMクラスタ（例：OM_C_1、OM_C_2、…OM_C_O）を生成し（⑤）、分析・評価機能に挿入する（⑥）。
- クローラ：スパム・メール内のURLに関連付けられたウェブサイトにアクセスし、そのHTMLコンテンツをダウンロードし（⑦）、文書に基づくクラスタリング機能に挿入する（⑧）。
- 文書に基づくクラスタリング機能：**3.5**で説明するように、ウェブ上の文書で確認した類似度に基づいて文書クラスタ（例：DC_C_1、DC_C_2、…DC_C_d）を生成する（⑨）。
- 分析・評価機能：O-means法によるクラスタリング機能と文書に基づくクラスタリング機能で得られる分析結果を使用して、O-means法によるクラスタリング手法と本論文が提案する12種類の統計に基づく特徴量に関するパフォーマンスを見積もり、スパムを送信するシステムとURLの参照先（ウェブ・サーバー）の分析を行う（⑥および⑩）。

3.2 ヘッダーに基づくクラスタリング機能

スパムを送信するシステム間のリレーションシップを見極めるために、「差出人」や「宛先」のような電子メールのヘッダーを活用する。

電子メールのヘッダーの特徴は電子メールを送信するために使用される電子メールのクライアント・プログラムに左右されるということである。

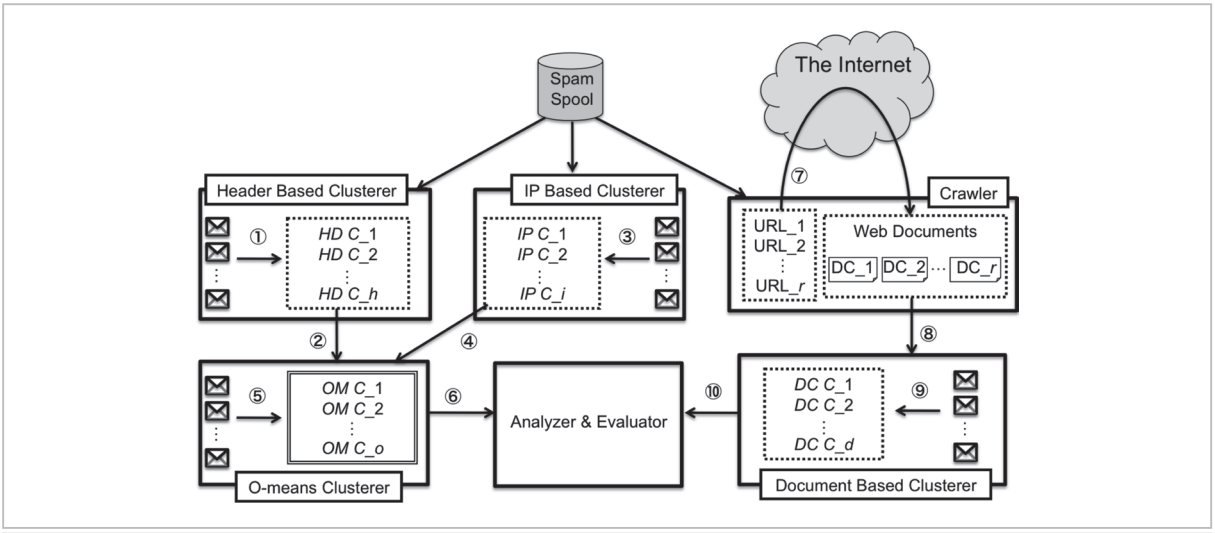

図1 O-means法によるクラスタリング手法のプロシージャの概要

つまり、多くの種類の電子メールのヘッダーが存在し、ヘッダーには重要なものもオプションとして使用されるものも存在するため、2つの電子メールのヘッダーが同一でない場合は、異なる電子メールのクライアント・プログラムによって送信されたものと考えることができる。さらに、多くの場合スパーマーと攻撃者は個別のスパム送信プログラムを使用しているものの、電子メールのヘッダーで確認した特徴を見極めることでプログラムを識別することができる。

ヘッダーに基づくクラスタリング機能においては、電子メールのヘッダーの種類、数、および順序の3つの基準に着目のうえ、スパム・メールをHDクラスタに分類する（同じ電子メールのヘッダーを持つスパム・メールが同じHDクラスタのメンバーになる）。図2は、本機能によるクラスタリングのプロセスを示している。クラスタリングのプロセスにおいて、本機能はまず各電子メールから電子メールのヘッダーを抽出し（①）、当該ヘッダーのタイプ（重複するタイプを除く）を抽出する（②）。その後、ヘッダーの数と順序が全く同じ場合は、2つの電子メールを同じHDクラスタとして認識する（③）。

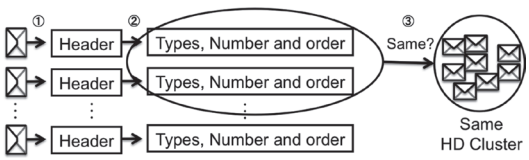

図2 ヘッダーに基づくクラスタリング機能のクラスタリングのプロセス

3.3 IPに基づくクラスタリング機能

図3は、文献[9]および[10]が提案するIPに基づくクラスタリングを示している。クラスタリングのプロセスにおいて、まず本機能によって各電子メールから取得したオリジナルのURLからユニークURLを抽出する（①および②）。その後ユニークURLからIPアドレスを抽出し（③）、ユニークURLから抽出したIPアドレスのセットが全く同一の場合は、2つの電子メールを同一のIPクラスタと認識する（④）。

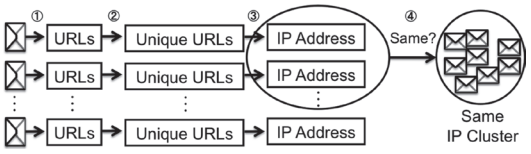

図3 IPに基づくクラスタリング機能のプロセス

3.4 O-means法によるクラスタリング機能

図4は、本論文が提案するO-means法によるクラスタリング手法のクラスタリングのプロセスについて示している。クラスタリングのプロセスにおいて、3.4.1で説明するとおり各電子メールから12種類の統計に基づく特徴量をまず抽出し（①）、3.4.2で説明する正規化手法に基づいて特徴量の値を正規化する（②）。その後、3.4.3で説明するとおりIPクラスタとHDクラスタを使用してK個の初期の中心を構築する（③、④、⑤、および⑥）。その後、K個の初期の中心とK-means法によるクラスタリング手法を使用して、スパム・メールからOMクラスタ（例：OM_C_1、OM_C_2、…OM_C_O）を構築する（⑥および⑦）。

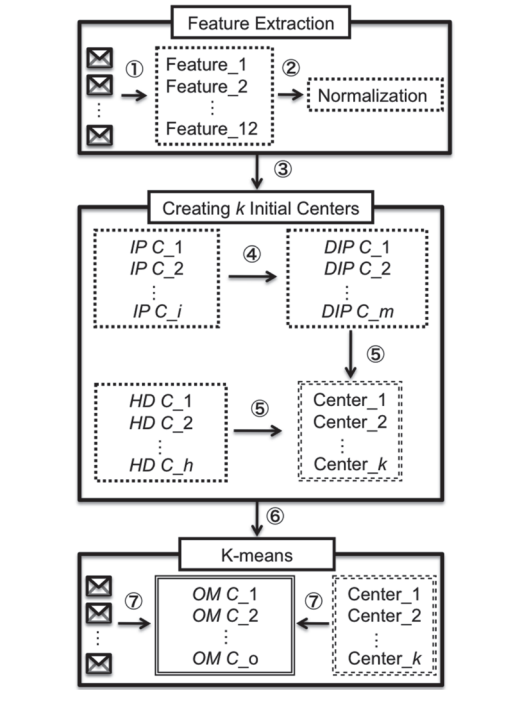

図4 O-means法によるクラスタリングのプロセス

3.4.1 特徴量の抽出

この特徴量の抽出は、電子メール、URL、およびドメイン名のコンテンツが頻繁に変更される場合でも、スパーマーは電子メール送信プログラムに組み込まれた一定のパターンに基づいてスパム・メールおよび特にURLを構築するという前提に基づいている[1][3][9][10][12]。したがって、スパム・メールやURLに関するルールを識別することによって各スパーマーを区別できる可能性がある。この目的のために、表1のとおり12種類の統計に基づく特徴量を定義し、図5で示される電子メールの例に基づいて当該特徴量について説明する。

図5では、電子メールには電子メールのヘッダーを示す7行の情報と本文を示す1つのURLが含まれている。本電子メールから電子メールのバイト数（310バイト）と行数（8行）をまず計算する。

その後、電子メールからユニークURLを抽出し、ドメイン名、パス、およびクエリーの3つの要素に分割する。さらに、クエリー部分には複数のキーと値のペアが含まれるため、クエリーをキーと値の各ペアに分割する。これらの要素に基づいて、表1で示されるとおりユニークURLに関連付けられた9種類の特徴量（番号3から番号11まで）の値を計算する。最後に、Alexa.comが提供するグローバル規模で最も使用されるトップ100のウェブサイトのURLの数を数える。本特徴量を使用する理由は、スパーマーが正当なURL（特にGoogle.comやYahoo.com等のよく使用されるURL）を含むスパム・メールを作成することで、URLに基づくスパム検出メカニズムを回避し混乱させようと考えることがあるためである（①）。したがって、よく使用されるURLの使用頻度を確認することで、スパーマーの特徴を見極めることができる場合がある。

3.4.2 正規化

スパム・メールから12種類の統計に基づく特徴量を抽出した後、各特徴量には異なるスケールが設定されているため特徴量の値の正規化を行う必要がある。本正規化手法は、主に文献[13]のとおりである。上記の12種類の統計に基づく値を持つ一連のデータ・インスタンス（スパム・メール）に基づいて、以下のとおり各インスタンスの正規化された値を計算する。

$$\text{normalized\_instnace}[r] = \frac{\text{original\_instance}[r] - \text{average}[r]}{\text{standard\_deviation}[r]} \quad (1)$$

$$\textbf{subject to :}\ \forall r\ (1 \leqq r \leqq 12)$$

上記において、[r]はr番目の特徴量を指し、average[r]およびstandard_deviation[r]はそれぞれ全てのデータ・インスタンスから得られた12種類の特徴量の平均と標準偏差を指す。本正規化手法によって、標準偏差に基づいて各データ・インスタンスの各特徴量値が該当する特徴量の平均からどれくらい乖離しているかを示す。

表1 12種類の統計に基づく特徴量

番号	特徴量の名称	値
1	電子メールのサイズ	310
2	行数	8
3	ユニークURLの数	1
4	ユニークURLの長さの平均	57
5	ドメイン名の長さの平均	17
6	パスの長さの平均	10
7	クエリーの長さの平均	22
8	キーと値のペアの数の平均	2
9	キーの長さの平均	5.5
10	値の長さの平均	4
11	ドメイン名内のドット(.)の数の平均	1
12	グローバルのトップ100のURLの数	0

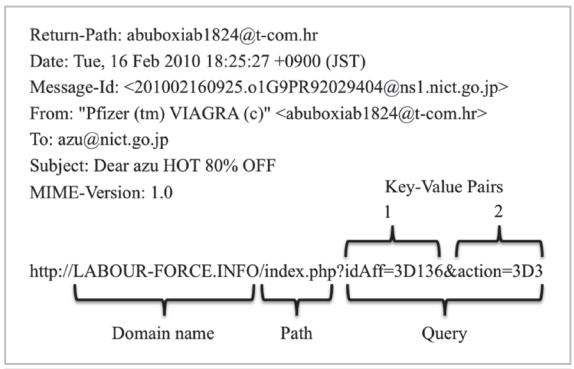

図5 電子メールの例

3.4.3 K個の初期の中心の作成

(1) 目的と戦略

スパム・メールから上記の12種類の正規化された値を取得後、次のクラスタリングのフェーズ(K-means法によるクラスタリング手法)で使用するK個の初期の中心を作成する。K-means法によるクラスタリング手法のパフォーマンスを最大化するために、特定のスパム・メール内で実際のK個のクラスタを代表するK個の初期の中心を選択する必要がある。スパム・メールについては、制御を行うエンティティ(例: ボットネット)から送信され、そのURLが同一のウェブ・ページに関連付けられている一連のスパム・メールとして各クラスタを定義することができる。つまり、2つの電子メールが異なるボットネットから送付されている場合は、同じウェブ・ページに関連付けられた同一のURLを共有している場合でも、異なるクラスタのメンバーとして認識される。

これらのクラスタをK個の初期の中心に反映するために、IPクラスタとHDクラスタを使用する。そのために、まずIPクラスタをDIPクラスタ(重複クラスタ、例: DIP_C_1、DIP_C_2、…DIP_C_m)にマージする。当該DIPクラスタのメンバー(スパム・メール)はユニークURLから抽出された1つ以上のIPアドレスを共有している。すなわち、各DIPクラスタのメンバーは参照先URLについてお互いに緊密なリレーションシップを保有している可能性が高くなる。しかしながら、IPクラスタが同じIPアドレスを持つ同じウェブ・サーバーのホスティングを行っている場合でも、制御を行う異なるエンティティ(当該エンティティのURLは異なる種類のウェブ・ページに関連付けられている)から送信された大量の関連性のない電子メールがIPクラスタには含まれている。したがって、HDクラスタに基づいてDIPクラスタをK個のグループに分割する。当該HDクラスタのメンバー(スパム・メール)は、電子メールのヘッダーのタイプ、ヘッダーの数、およびヘッダーの順序の点で同じ処理を行うスパム送信システムから送信される。その結果、個別のスパマーが同じウェブ・ページに関連付けられたスパム・メールを送信した場合でも、本手法は各電子メールを1つのグループとして検出することができる。さらに、あるスパマーが異なるウェブ・サーバー(IPアドレス)上でホスティングされた様々な種類の複数のウェブ・ページに関与している場合でも、本手法を活用することによる個別のグループとして明確に分類することができる。

(2) DIPクラスタの作成

図6はDIPクラスタを作成するマージのプロセスについて説明している。本プロセスは以下の7つのステップで実行される。

① 全てのIPクラスタを一連の中間IPクラスタにフィードする。
② 一連の中間IPクラスタから最も大きなIPクラスタ($IP\_C_{largest}$)を選択する。
③ 初期のメンバーとして$IP_C_{largest}$のみを含む新規のDIPクラスタを作成する。
④ 一連の中間IPクラスタから$IP_C_{largest}$を除く全てのIPクラスタを選択する。
⑤ 全てのIPクラスタを2種類に分類する。あるIPクラスタが1つ以上の同一のIPアドレスを$IP_C_{largest}$と共有している場合は、新規のDIPクラスタのメンバーとなる。そうでない場合は、一連の中間IPクラスタに戻る。
⑥ 新規のDIPクラスタを一連のDIPクラスタに追加する。
⑦ 一連の中間IPクラスタが空になるまで②から⑦のプロセスを繰り返す。中間IPクラスタが空にならない場合は、マージのプロセスは終了する。

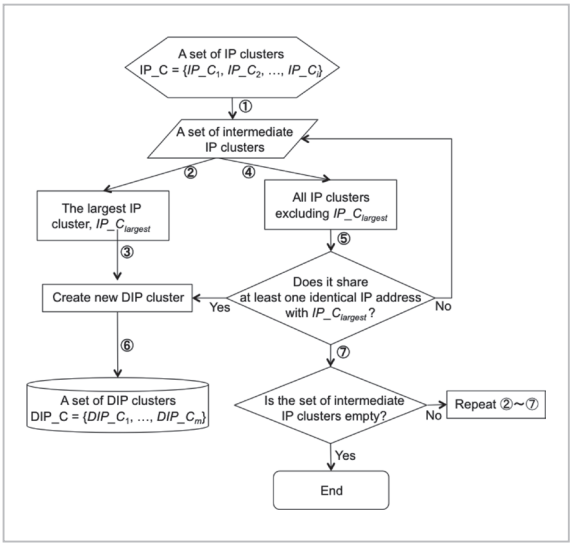

図6 DIPクラスタの作成にあたってのマージ・プロセス

この結果、一連のDIPクラスタ（DIP_C_1、DIP_C_2、…DIP_C_m）を取得することができる。

(3) K個のグループの作成

図7はK個の初期の中心の作成プロセスを説明している。本プロセスは以下の5つのステップで実行される。

① 全てのDIPクラスタを一連の中間DIPクラスタにフィードする。
② 一連の中間DIPクラスタからあるDIPクラスタを選択する。
③ 当該DIPクラスタのメンバー（スパム・メール）をグループに分割する。本グループにおいては、同じグループ内のスパム・メールが同じHDクラスタに属する。
④ 一連のK個のグループに新規のグループを追加する。
⑤ 中間DIPクラスタが空にならない場合はステップ②に戻り、中間DIPクラスタが空になった場合は本作成プロセスは終了する。

上記の作成プロセスに基づいて、K個のグループ（g_1、g_2、…g_k）が作成される。K個のグループから、本プロセスによってメンバーの平均値が計算され、K個の初期の中心として認識される。

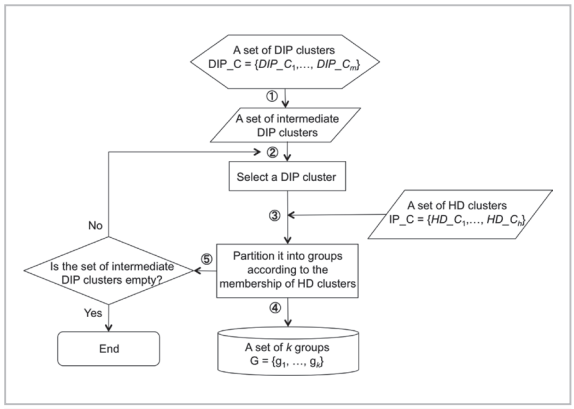

図7 K個の初期の中心の作成プロセス

K個の初期の中心を作成するにあたり、スパム・メールに対してK-means法によるクラスタリング手法を適用する。K-means法によるクラスタリング手法のクラスタリングの実施の結果、OMクラスタ（OM_C_1、OM_C_2、…OM_C_O）を取得することができる。この実験では、ユークリッド距離を使用して2つの電子メールの間の距離を計算していることに注意する必要がある。実際のd次元の空間R^dにおけるベクトルであるオブジェクトのペア（例：$\mathbf{a}=\{a_1,a_2,\ldots,a_d\}$および$\mathbf{b}=\{b_1,b_2,\ldots,b_d\}$）に基づいて、$\mathbf{a}$と$\mathbf{b}$の間のユークリッド距離（$d(\mathbf{a},\mathbf{b})$）は以下のとおり計算される。

$$d(\mathbf{a},\mathbf{b}) = \sqrt{\sum_{i=1}^{d}(a_i - b_i)^2}$$

3.5 文書に基づくクラスタリング機能

本クラスタリング手法を検証するために、URLに関連付けられたウェブ・ページで検出した類似度に基づいてスパム・メールを文書クラスタ（DC_C_1、DC_C_2、…DC_C_d）にクラスタリングした。

ウェブ・ページ間の類似度を検証するために、該当するウェブ・ページに対してText shingling[1][10][14]を適用した。図8はText shinglingの例を説明しており、AとBの2つのWeb上の文書には50語（A_1、A_2、…A_{50}）と40語（B_1、B_2、…B_{40}）が含まれ、1つのshingleのサイズは5である。1つのshingleのサイズが5であるため、AとBからそれぞれ46のshingle（例：$A_1A_2A_3A_4A_5$、$A_2A_3A_4A_5A_6$）と36のshingle（例：$B_1B_2B_3B_4B_5$、$B_2B_3B_4B_5B_6$）を構築し、その後AとBの間の類似度を以下のとおり計算する。

$$\frac{\text{同一のshingleの数}}{\text{AとBに含まれるユニークなshingleの数}}$$

この例では、AとBの間で32のshingleが共通しており、この2つが同一のクラスタであるかど

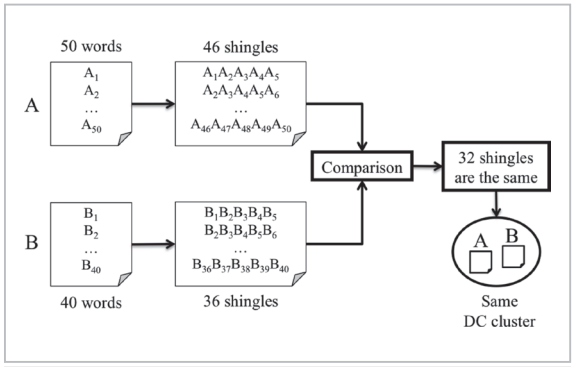

図8 Text shinglingの例

うかを決定する閾値が50%である場合、この2つの文書は同一のクラスタであるとみなす。AとBの間の類似度が64%={32÷(14+32+4)}であり、50%を超えているためである。

4　実験の結果

4.1　ダブルバウンスメール

ダブルバウンスメールでは、ヘッダーのスパムの差出人や宛先に関連付けられた有効な電子メールアドレスは存在しない。図9はダブルバウンスメールのプロセス全般を示している。a@user.com、b@user.com、c@user.comの電子メールを使用する3名の有効なユーザーが存在するとする。あるスパーマーがリターンパスのアドレス（forged@address.com）と宛先のアドレス（unknown@address.com）が実際には存在していないダブルバウンスメールを対象となるSMTPサーバーに対して送信すると（①）、「ユーザーが存在しません」という旨のエラー・メッセージが2つのSMTPサーバー間で交換される（②および③）。この場合、スパーマーは活動を隠蔽するためにリターンパスのアドレスを意図的に偽造し、実際には存在しない宛先のアドレスをランダムに生成したと言える。

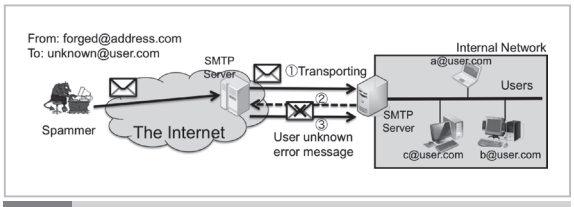

図9　ダブルバウンスメールのプロセスの概要

しかしながら、通常の場合では、送信者が電子メールの宛先のアドレスを間違って記載した場合でも、電子メールのヘッダーには1つ以上の有効なリターンパスのアドレスが含まれている。この状況においては、ダブルバウンスメールは純粋なスパムと認識することができるため、本実験ではダブルバウンスメールを分析データ（スパム・メール）として使用する。

4.2　実験データの説明

本実験では3週間（2010年1月25日から2月20日まで）において作者のSMTPサーバーに到着した596,526件のダブルバウンスメールを収集した。図10は、当該メールのプロパティの概要について説明している。全ての電子メールのうち、526,544件の電子メールには本文中に1つ以上のURLが含まれており、URLの総数は1,405,950個であり、そのうちダウンロード可能なURLとユニークURLの数はそれぞれ1,296,120個と1,048,158個であった。また、275,117個のユニークなIPアドレスを使用して全てのスパムを送付していることが分かった。これにより、ユニークなIPアドレスのそれぞれが平均でおよそ1.9件の電子メールと関連していることが分かる。さらに、URLの参照先と関連付けられているユニークなIPアドレスの総数は1,643個であった。

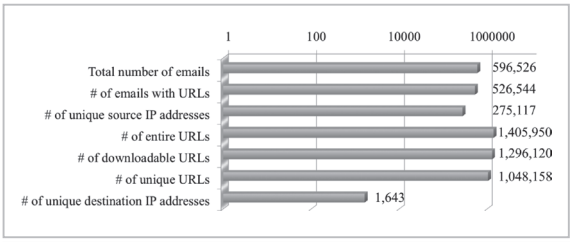

図10　実験データのプロパティの概要

図11は、スパムを送信するシステムと参照先のURLの国別分布を示している。図11によると、スパム・メールの10%と8%がそれぞれアメリカとブラジルから送信されているものの、今回使用したデータによると全スパムは合計204か国から送信されていることが分かる。一方、参照先の

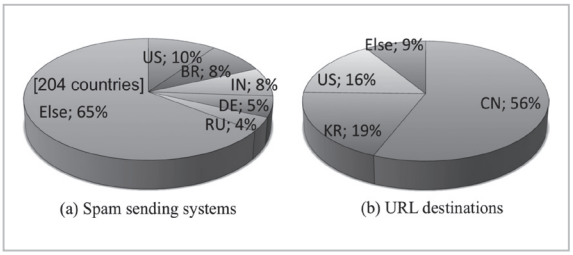

図11　スパムを送信するシステムと参照先のURLの国別分布

URL（ウェブ・サーバー）については、56%が中国に存在するものの、合計35%のウェブ・サーバーが韓国（19%）とアメリカ（16%）に存在している。これらの結果を検証すると、スパムを送信するシステムとウェブ・サーバーの地理的分布は大きく異なることが言える。すなわち、スパムを送信するシステムは世界中に広範に分散しているものの、ウェブ・サーバーは3か国のみに集中している。

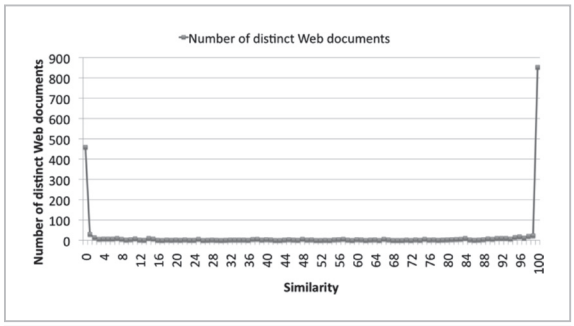

図12 1,739件のウェブ上の文書の類似度分布

4.3 文書に基づくクラスタリングの分析結果

O-means法によるクラスタリング手法のパフォーマンスを検証するために、**3.5**で説明したとおりダブルバウンスメールをDCクラスタにクラスタリングした。クラスタリングのプロセスにおいて、1つのshingleのサイズを5に設定し、閾値を50%に設定した。すなわち、あるshingleは5つの隣接した語で構成され、ウェブ上の文書間の類似度が50%を超えている場合は、同一のDCクラスタのメンバーとして認識される。今回の実験においては、1,296,120個のURLからクロールされた全てのウェブ上の文書のうち、1,739件の個別のウェブ上の文書についてハッシュ値が互いに異なることを確認した。

まず上記2種類の条件に基づいて1,739件のウェブ上の文書を分類し、772のグループ（独立したコンテンツを持つ772件のウェブ上の文書）を検出した。さらに、1,739件のウェブ上の文書のうち656件のウェブ上の文書が他の文書と類似度を持たず、1,083件のウェブ上の文書が他の文書と類似度を持っていることを確認した。さらに研究を進めると、772のグループのうち、最大規模のグループには156件の同様のウェブ上の文書がメンバーとして含まれていることを確認した。図12は、1,739件のウェブ上の文書を772のグループのうちの1つに割り当てた場合の類似度の分布について説明したものである。図12によると、ほとんどのウェブ上の文書の類似度は0か100のいずれかに近いことが分かる。すなわち、閾値（50%）が文書に基づくクラスタリング機能のパフォーマンスに影響を与えないということを示している。

その後、772件のウェブ上の文書のメンバー情報に基づいてダブルバウンスメールを分類した。すなわち、2つのダブルバウンスメールが772件のウェブ上の文書の1つにリンクされている1つ以上のURL（本URLは同一のものでない場合が多い）を共有している場合、当該メールを同一のDCクラスタのメンバーであると認識する。その結果、実験データによって772のDCクラスタを取得し、本クラスタにおいては、同一のDCクラスタ内のダブルバウンスメールがウェブ・ページが当該ページに含まれるURLに接続するという点で互いに類似度を持っていることを確認した。

4.4 O-means法によるクラスタリング結果

URLを含む526,544件の電子メールについて、それぞれ**3.2**および**3.3**で説明したとおりHDクラスタとIPクラスタを構築した。HDクラスタとIPクラスタの数は、それぞれ1,062個と1,204個であった。

さらに、**3.4.3**で説明したとおり1,204個のIPクラスタをDIPクラスタにマージし、900個のDIPクラスタを取得した。900個のDIPクラスタと1,062個のHDクラスタを活用のうえ、**3.4.3**で説明した作成プロセスに基づいて3,528個のグループを作成し、その結果メンバーの平均値を持つ3,528個の初期の中心の計算を行った。その後、3,528個の初期の中心をK-means法によるクラスタリング手法にフィードし、OMクラスタを作成した。最終的に、2,049個のOMクラスタを取得することができた。

2,049個のOMクラスタの確度を見積もるために、**4.3**で取得したDCクラスタを使用して確度を測定した。見積もりを行うにあたっては、最初にあるOMクラスタから1件以上の電子メールを含む一連のDCクラスタを取り出し、その後当該OMクラスタ内で最も多くの電子メールを共有する特定のDCクラスタを選択する。最終的に、以

下のとおりOMクラスタの確度を計算する。

$$\frac{\text{電子メールの最大数}}{\text{当該OMクラスタ内の電子メールの総数}}$$

検証を行った結果、2,049個のOMクラスタの平均の確度は86.63%であり、この数値はIPに基づくクラスタリング機能の確度（80.98%）よりもおよそ6%高くなった[10]。図13は、トップ100のクラスタに関するIPに基づくクラスタリング機能とO-means法によるクラスタリング機能の確度を示したものである。当該トップ100のクラスタは、本実験で使用したダブルバウンスメールの規模（メンバーの数）においてそれぞれのクラスタリング機能について98%と86%を占めている。特に、トップ12のクラスタのうち10のクラスタにおいて、O-means法によるクラスタリング手法はIPに基づくクラスタリング手法よりも優れたパフォーマンスを実現することを確認した。トップ10のOMクラスタにおける確度は、表2のとおりである。

表2は、トップ10のOMクラスタにおけるスパムを送信するシステムと参照先のURLの統計情報を示したものである。表2で示されているとおり、各OMクラスタには数多くの個別のスパムを送信するシステム（4,730個から27,476個）が含まれており、これは互いに緊密に連携するボットネットまたはシステムの規模を示している。さらに、当該システムは多くの国（90か国から151か国）に分散している。一方、参照先のURLについては、ユニークなIPアドレスの数は2個から5個であり、1か国から3か国に存在していること

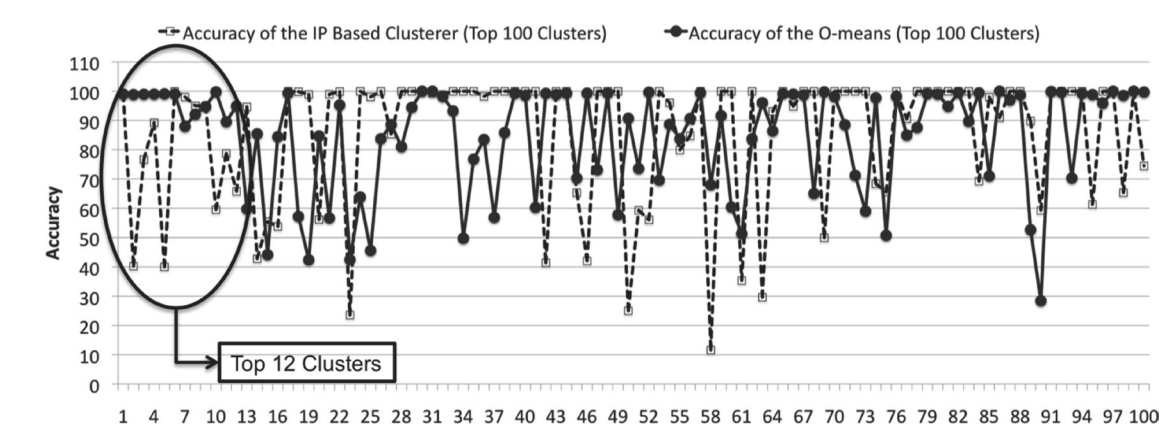

図13 トップ100のクラスタに関するIPに基づくクラスタリング手法とO-means法によるクラスタリング手法のパフォーマンス比較

表2 トップ10のOMクラスタの統計情報

	OMクラスタのID（トップ10のクラスタ）									
	1079	82	24	12	19	22	16	10	129	1193
確度	99.03%	98.84%	98.95%	98.98%	99.08%	99.05%	87.99%	92.12%	94.74%	99.76%
電子メールの数	31,790	22,857	20,230	17,817	9,850	9,409	9,369	9,013	8,998	6,772
ユニークな配信元IPアドレスの数	27,476	20,176	18,244	16,260	9,229	8,901	7,681	7,368	5,731	4,730
ユニークな配信元の国の数	108	108	106	106	92	90	94	91	149	151
URLの総数	**95,370**	**68,619**	**60,688**	**53,447**	**29,548**	**28,225**	**56,214**	**54,073**	8,998	6,772
ユニークURLの数	**94,805**	**68,198**	**60,294**	**53,151**	**29,532**	**28,068**	**56,008**	**53,958**	12	5
ユニークなドメイン名の数	**94,805**	**68,198**	**60,294**	**53,151**	**29,532**	**28,068**	**56,008**	**53,958**	12	5
ユニークな参照先IPアドレスの数	3	3	3	3	3	3	3	5	5	2
ユニークな参照先の国の数	1	1	1	1	1	1	1	2	3	2

が分かる。実際に本研究においては、当該URLは中国、韓国、アメリカの3か国のいずれかに存在していることを確認した。表2で最も重要なポイントは、各OMクラスタおよび特にトップ8のOMクラスタにおいては多くの個別のURLとドメイン名が存在していることである。これによって、スパーマーや攻撃者はURLやドメイン名を頻繁に変更することが分かる。したがって、既存のドメイン名やURLに基づくクラスタリング手法の場合は、ほとんどのURLがユニークなドメイン名を持つため、クラスタリングの確度が非常に低くなると言える。

4.5 特徴量選択手法

本セクションではスパム・メールのヒューリスティック分析に基づく特徴量選択手法について説明する。

本手法では、まず**3.5**で説明した文書に基づくクラスタリングで取得したオリジナルの772個のクラスタからトップ10のクラスタを選択した。表3はトップ10のクラスタに関する統計情報について説明している（詳細は**4.5.1**を参照）。トップ10のクラスタを使用して、12種類の特徴量をヒューリスティックに評価し、スパム・メールのクラスタリング確度を向上するための貢献度について検証した（**4.5.2**を参照）。

4.5.1 トップ10のクラスタ

3.5で説明した文書に基づくクラスタリングを使用して772個のクラスタを取得し、本研究で使用したSMTPサーバーに送信されるURLを含むスパム・メールのおよそ90%の原因となるトップ10のクラスタについて確認した。表3は当該クラスタの統計情報を示している。表3によって、各クラスタには数多くの個別のスパムを送信するシステム（3,061個から169,149個）が含まれており、これは互いに緊密に連携するボットネットまたはシステムの規模を示している。さらに、当該システムは多くの国（101か国から192か国）に分散している。一方参照先のURLについては、ユニークなIPアドレスが3個から164個にとどまり、2か国から29か国に存在していることが分かる。また、URLとドメイン名には2つのパターンがあることが分かる。1、3、および5のクラスタについては、スパーマーや攻撃者はURLとドメイン名を頻繁に変更するものの、他のクラスタについては同一のURLとドメイン名をほぼ継続的に使用している。

4.5.2 重要な特徴量の選択

スパム・メール間の類似度を計算するために12種類の統計に基づく特徴量の定義を行い、クラスタリングの確度が既存の方法よりも優れていることを示したものの、12種類の統計に基づく特徴量は互いに独立して存在しているわけではなく、クラスタリングの確度の向上には役立たない冗長な特徴量がいくつか存在する可能性があることを検証する必要がある。さらに、大量のスパム・メー

表3 トップ10のクラスタの統計情報

	トップ10のクラスタ									
ID	1	2	3	4	5	6	7	8	9	10
A	286,024	30,205	22,696	21,061	17,393	10,741	9,807	7,649	5,601	5,265
B	169,149	19,102	16,881	10,480	4,003	4,828	6,101	5,241	4,958	3,061
C	192	177	163	162	101	120	120	127	133	137
D	951,565	30,209	77,794	21,061	17,397	18,363	9,806	7,649	5,919	5,254
E	891,795	177	68,584	94	16,790	6,306	102	67	155	8
F	890,022	157	68,344	79	16,757	6,300	97	63	135	6
G	80	11	164	14	100	21	3	3	13	3
H	11	6	29	5	30	4	2	3	4	3

A：電子メールの数、B：ユニークな配信元IPアドレスの数、C：ユニークな配信元の国の数、D：URLの総数、E：ユニークURLの数、F：ユニークなドメイン名の数、G：ユニークな参照先のIPアドレスの数、H：ユニークな参照先の国の数

ルに対応したうえで効果的な分析を行う必要があることを考慮すると、オリジナルの12の特徴量から重要な特徴量を抽出する必要がある。それによって、クラスタリング確度を高く保持したうえでスパム・メールの分析時間を短縮できるようになるためである。

一連の最適化された特徴量を発見するには、12種類の特徴量の全ての組み合わせを検証することが最も効果的である。全てのケースの数は$\sum_{i=1}^{12} {}_{12}C_i$として見積もることができるが、そのような計算を行うには非常に時間がかかる。したがって、トップ10のクラスタに関する以下のヒューリスティック分析に基づいて別の実験を行った。

1. 各クラスタについて、同じ値を持つスパム・メールの数は12種類の特徴量の全てに関して計算される。
2. トップ10のクラスタのサイズはお互いに異なるため、各クラスタのサイズに基づいてスパム・メールの数は0または1のスケールで計算する。
3. 各特徴量に関して、図14のとおり横軸が各特徴量の値を示し、縦軸が各クラスタ内のスパム・メールの数の比率を示す2次元のグラフを作成した。
4. 図14の分析結果を使用して、「電子メールのサイズ（14（a））」、「行数（14（b））」、「URLの長さ（14（d））」の3種類の重要な特徴量をまず選択した。ほとんどのクラスタはこれらの値を使用することで互いに区別することができるためである。つまり、各クラスタ内のスパム・メー

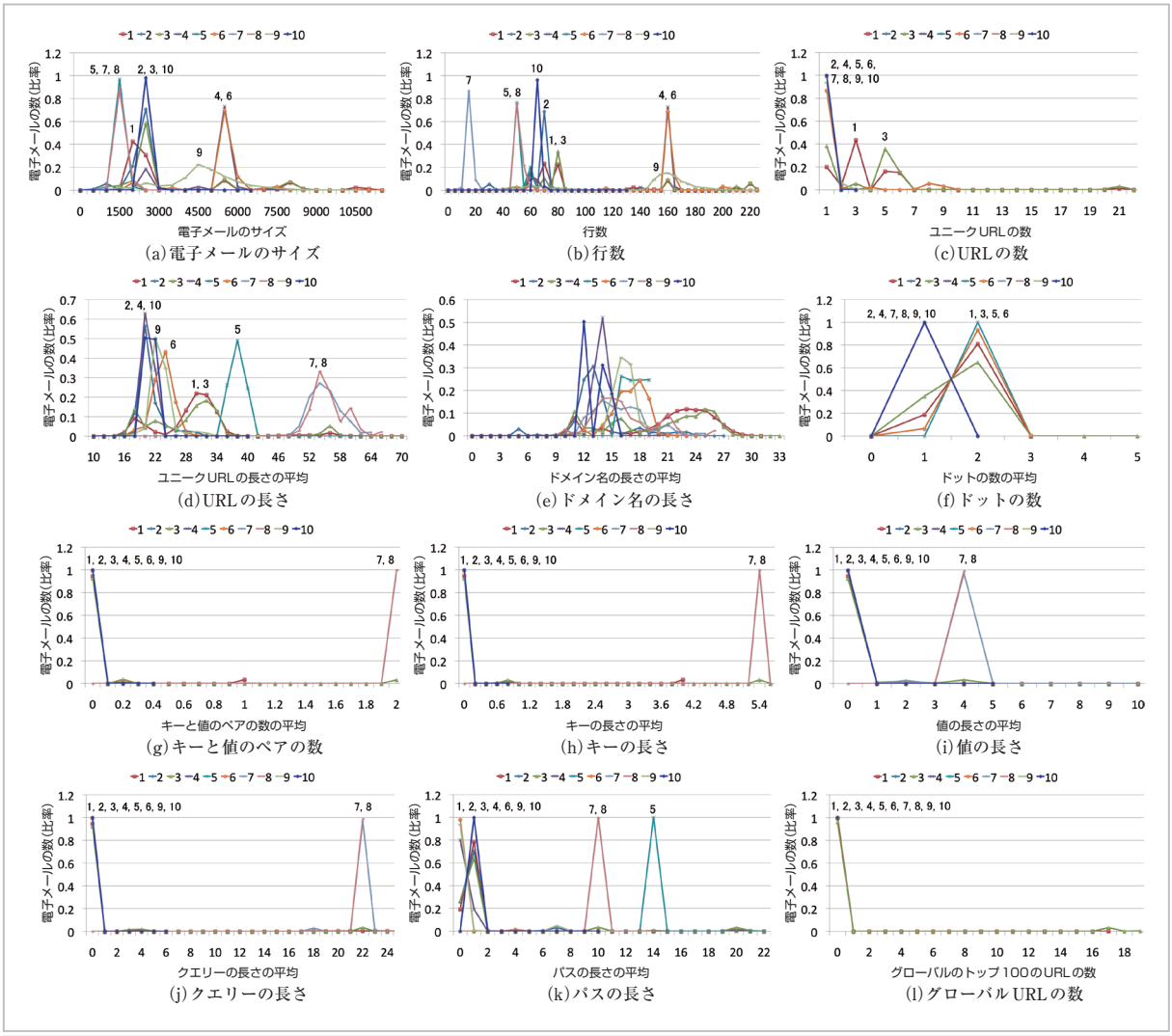

図14　トップ10のクラスタに関する12種類の統計に基づく特徴量の値の分散

ルはこれら3種類の特徴量に関して値の分布が異なるため、スパム・メール間の類似度を計算するためにこれら3種類の特徴量を使用すれば、同一クラスタ内のメンバーが同様の値を持つようなクラスタにスパム・メールを正確に分類することが可能になる。

5. 第2に、重要な特徴量の一覧から「グローバルURLの数(14(l))」を排除した。トップ10のクラスタはこの特徴量に関しては同様の値を持っているためである。
6. 第3に、「ドメイン名の長さ(14(e))」は重要な特徴量と考えなかった。この値の分布が他のクラスタとは異なるクラスタが存在せず、各クラスタの区別には全く役に立たないためである。
7. 第4に、「URLの数(14(c))」、「キーと値のペアの数(14(g))」、「キーの長さ(14(h))」、「値の長さ(14(i))」、「クエリーの長さ(14(j))」、および「パスの長さ(14(k))」の6種類の特徴量は重要な特徴量の一覧から除外した。これらは7、8、1、3、および5のクラスタを他のクラスタと区別するためには役立つものの、ステップ4で示される3種類の重要な特徴量を使用してもこれらのクラスタの区別を行うことができることが明らかであるためである。
8. 最終的に、ステップ4で示される3種類の重要な特徴量を使用することでクラスタ4とクラスタ5を区別することはできないため、この2つのクラスタを区別することができる「ドットの数(14(f))」を重要な特徴量の一覧に加えた。

この結果、「電子メールのサイズ」、「行数」、「URLの数」、および「ドットの数」の4種類の重要な特徴量を選択した。本手法はトップ10のクラスタのラベル情報を使用するため、管理された特徴量選択であることに注意する必要がある。

4.5.3 パフォーマンスの評価

4種類の重要な特徴量の効果を実証するために、当該特徴量のクラスタリングの確度と計算時間の検証を行った。クラスタリングのアルゴリズムとして、O-means法クラスタリング手法を採用した。検証結果は表4のとおりである。表4が示すとおり、この4種類の重要な特徴量を使用するだけでほぼ同じクラスタリングの確度(86.33%)を実現できることが分かる。さらに、この4種類の重要な特徴量を使用するだけで実行時間を劇的に削減することができ、スパム・メールをより効果的に分析できるようになることも確認した。検証は3GBのRAMを搭載した2.8GHzのIntel Core 2 Duoプロセッサのマシンで行い、プログラムはPerlのプログラミング言語とMysqlで作成したものである。

表4 クラスタリングの確度と計算時間の比較

	4種類の重要な特徴量	全ての特徴量
クラスタリングの確度	86.33%	86.63%
実行時間(秒)	6,124	28,772

5 考察

O-means法によるクラスタリング手法のパフォーマンスを評価するために、**3.5**で説明したText shingling手法を使用して文書クラスタ(例: DC_C_1、DC_C_2、…DC_C_d)を構築した。テキスト文書の類似度を見積もるには多くの手法(例: レーベンシュタイン距離)が存在するものの、本手法ではText shinglingの手法を活用してURLからダウンロードしたウェブ・ページの類似度を比較した。本手法はウェブ・ページの類似度を測定および比較するために開発され、その効果が多くのアプローチで検証済みであるためである[1][3][10][14][16][17]。さらに、図12で示される実験結果のとおり、1,739件のウェブ上の文書はこのText shinglingの手法を活用することで当該文書間の類似度に基づいて適切に分類を行うことができた。

本実験では、作者のSMTPサーバーに送信される3週間分のダブルバウンスメールを使用してO-means法によるクラスタリング手法の評価を行った。より正確にO-means法によるクラスタリング手法の評価を行うためには、より長期間にわたるダブルバウンスメールやURLに関連づけられたウェブ・ページを分析したほうが効果的である。しかし、現実的な問題によってこれを行うことが

困難になっている。ウェブ・クローリングによる侵入プロセスによって、スパマーが特定のグループのユーザーはスパム・メールの攻撃を受けやすいことに気づき、当該ユーザーにより多くのスパムを送信するようになることが考えられるためである[1]。その結果、URLからダウンロードされた全てのウェブ上の文書間の類似性を計算する時間が飛躍的に増大する。したがって、本実験と同様に既存のアプローチのほとんどでは、短期間に送信されたスパム・メールを使用して各システムの検証を行っている。文献[1][7][8]においては、それぞれ1か月分、1週間分、および9日間分に限られたスパム・メールを使用している。その条件があるものの、これらのアプローチではいくつかのボットネットやスパム・クラスタの検出に成功している。このような結果の理由の1つとしては、最近のボットネットが効力を発揮する時間や単一のスパム・メッセージの有効期間が非常に短いことが挙げられる[2][3]。これを考慮すると、本実験は妥当な形態で実施されたものと結論付けることができる。

6　結論

本論文では、最も広範に使用されているクラスタリング手法の1つであるK-means法のクラスタリング手法[11]に基づいて、最適化されたスパム・クラスタリング手法であるO-means法のクラスタリング手法を提案している。O-means法のクラスタリング手法は、K-means法のクラスタリング手法の欠点を克服することでパフォーマンスを向上する。K-means法によるクラスタリングの結果は、選択されたK個の初期の中心に左右され、適切な数（K個）のクラスタを事前に設定することが非常に困難であることが分かっている。作者のSMTPサーバーで収集された3週間分のスパムを検証することで、O-means法によるクラスタリング手法の確度はおよそ87%であり、この値は以前のクラスタリング手法よりも優れていることを確認した。さらに、スパム・メール間の類似度を比較するために、新たに12種類の統計に基づく特徴量の定義を行い、O-means法によるクラスタリング手法の効率を向上するために役立つ、最適化された一連の特徴量を見極めるための特徴量選択手法について提案した。本手法を活用することで、計算時間を短縮したうえで86.33%のクラスタリング確度を実現する4種類の重要な特徴量の特定を行うことができた。

本論文ではクラスタリングの確度を向上することを主に説明したが、O-means法によるクラスタリング手法で取得できるスパム・クラスタを使用することで、より正確にスパムに基づく攻撃を分析できるようになる。したがって、将来においてはスパムに基づく攻撃についてより現実的な分析を実施することによって、スパムを送信するシステムと悪意のあるウェブ・サーバーのインフラを検出し、それらのシステムがどのように連携しているかを確認できるようにする必要がある。さらに、スパム・クラスタを活用することでどれくらいURLに関連付けられたウェブ・ページの分析時間の削減が可能になるかを調査することも重要である。O-means法によるクラスタリング手法が生成したOMクラスタのマージを検討する必要もある。K個の初期のクラスタの作成プロセスにおいて、2つのスパム・メールが異なる制御を行うエンティティから送付されている場合は、同じウェブ・ページに関連付けられた同一のURLを共有している場合でも、異なるクラスタのメンバーとして認識するためである。類似度に基づいてOMクラスタのマージを行うことによって、OMクラスタ間の関係性を見極めることができるようになり、同じウェブ・ページに関連付けられているURLを共有する電子メールを含むOMクラスタのグループを特定できるようになると思われる。追加の統計に基づく特徴量を構築し、特徴量をあらゆる組み合わせで検証することで、スパム・メールのクラスタリングに役立つ、最適化された一連の統計に基づく特徴量を入手できるようにすることは、大きな課題でもある。最後に、より効果的にスパムに基づく攻撃を分析するために、様々なドメインソースから取得できるより大容量のスパムのデータセットに基づいてO-means法によるクラスタリング手法の検証を行うことを計画している。

参考文献

1 Xie, Y., Yu, F., Achan, K., Panigrahy, R., Hulten, G., and Osipkov, I., "Spamming botnets: signatures and characteristics," ACM SIGCOMM Computer Communication Review, Vol. 38, No. 4, October 2008.

2 Ramachandran, A. and Feamster, N., "Understanding the network-level behavior of spammers," Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications, Vol. 36, No. 4, September 11–15, Pisa, Italy, 2006.

3 Anderson, D. S., Fleizach, C., Savage, S., Voelker, and G. M., "Spamscatter: Characterizing Internet Scam Hosting Infrastructure," Proceedings of the USENIX Security Symposium, Boston, 2007.

4 Jennings, R., "The global economic impact of spam, 2005 report," Technical report, Ferris Research, 2005.

5 Kawakoya, Y., Akiyama, M., Aoki, K., Itoh M., and Takakura, H., "Investigation of Spam Mail DrivenWeb-Based Passive Attack," IEICE Technical Report, ICSS2009-5(2009-5), pp. 21–26, 2009.

6 Jungsuk Song, Daisuke Inoue, Masashi Eto, Suzuki Mio, Satoshi Hayashi, and Koji Nakao, "A Methodology for Analyzing Overall Flow of Spam-based Attacks," 16th International Conference on Neural Information Processing(ICONIP 2009), LNCS 5864, pp. 556–564, Bangkok, Thailand, 1–5 December 2009.

7 Li, F. and Hsieh, M.H., "An Empirical Study of Clustering Behavior of Spammers and Group-based Anti-Spam Strategies," Proceedings of 3rd Conference on Email and Anti-Spam (CEAS), pp. 21–28, Mountain View, CA, 2006.

8 Zhuang, L., Dunagan, J., Simon, D. R., Wang, H. J., and Tygar, J. D., "Characterizing botnets from email spam records," Proceedings of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats, No. 2, pp. 1–9, San Francisco, 2008.

9 Jungsuk Song, Daisuke Inoue, Masashi Eto, Hyung Chan Kim, and Koji Nakao, "Preliminary Investigation for Analyzing Network Incidents Caused by Spam," The Symposium on Cryptography and Information Security (SCIS 2010), Takamatsu, Japan, 19–22 January, 2010.

10 Jungsuk Song, Daisuke Inoue, Masashi Eto, Hyung Chan Kim, and Koji Nakao, "An Empirical Study of Spam : Analyzing Spam Sending Systems and Malicious Web Servers," 10th Annual International Symposium on Applications and the Internet (SAINT 2010), Seoul, Korea, 19–23 July, 2010.

11 MCQUEEN, J, "Some methods for classification and analysis of multivariate observations," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297, 1967.

12 John, J. P., Moshchuk, A., Gribble, S. D., and Krishnamurthy, A., "Studying spamming botnets using Botlab," Proceedings of the 6th USENIX symposium on Networked systems design and implementation, pp. 291–306, Boston, April 22–24, 2009.

13 L. Portnoy, E. Eskin, and S. Stolfo, "Intrusion Detection with Unlabeled Data Using Clustering," Proceedings of ACM CSS Workshop on Data Mining Applied to Security, 2001.

14 D. Fetterly, M. Manasse, M. Najork, and J. L.Wiener, "A large-scale study of the evolution of web pages," Softw. Pract. Exper., 34(2), 2004.

15 Huan Liu and Lei Yu, "Toward integrating feature selection algorithms for classification and clustering," IEEE Transactions on Knowledge and Data Engineering, Vol. 17, pp. 491–502, 2005.

16 A. Broder, " On the Resemblance and Containment of Documents," Proceedings of the Compression and Complexity of Sequences (SEQUENCES ' 97), pp. 21–29, June 11–13, 1997.

17 N. SHIVAKUMAR and H. GARCIA-MOLINA, "Finding near-replicas of documents and servers on the web," Proceedings of the First International Workshop on the Web and Databases (WebDB' 98), pp. 204–212, March, 1998.

（平成 23 年 6 月 15 日 採録）

宋　中錫（Jungsuk Song）
ネットワークセキュリティ研究所
サイバーセキュリティ研究室研究員
博士（情報学）
ネットワークセキュリティ、スパム解析、IPv6 セキュリティ等