キーワードの時系列特性を利用した スパムブログの収集・類型化・データセット作成

佐藤　有記[†]　宇津呂武仁[†]　福原　知宏[††]　河田　容英[†††]　村上　嘉陽[†††]

中川　裕志[††††]　神門　典子[†††††]

† 筑波大学大学院システム情報工学研究科　〒 305-8573 茨城県つくば市天王台 1-1-1
†† 東京大学 人工物工学研究センター　〒 277-8568 千葉県柏市柏の葉 5-1-5
††† (株) ナビックス　〒 141-0031 東京都品川区西五反田 8-3-6
†††† 東京大学 情報基盤センター　〒 113-0033 東京都文京区本郷 7-3-1
††††† 国立情報学研究所　〒 101-8430　東京都千代田区一ツ橋 2-1-2

あらまし　本研究では，ブログにおいて検索頻度の高いキーワードを主として狙ったスパムブログの問題についての分析を行うことを主目的として，スパムブログデータセットを作成し，スパムブログの分析を進めている．スパムブログデータセットの作成においては，まず，キーワードによって検索されるブログサイトの生起数の推移を観測することによってバースト現象を確認し，バースト日において特に一日の投稿記事数の多いブログサイトを中心にブログサイトの収集を行う．次に，これらのブログサイトに対してスパム・非スパムの識別作業を行うともに，スパムブログに対して，アフィリエイトサイトへのリンクの形態や，ブログ本文中の文書のコピー元の分類，コピーの際の文書収集手順の分類等のいくつかの観点からスパムブログの類型化を行う．また，同一のスパマーが作成していると思われるスパムブログに対するスパマーの識別結果を付与する．以上の情報を考慮して，スパムブログデータセットを作成する．

キーワード　ブログ，スパムブログ，アフィリエイト

Collecting/Classifying Splogs and Developing Splog Data Set based on Time Series Characteristics of Keywords

Yuuki SATO[†], Takehito UTSURO[†], Tomohiro FUKUHARA[††], Yasuhide KAWADA[†††],

Yoshiaki MURAKAMI[†††], Hiroshi NAKAGAWA[††††], and Noriko KANDO[†††††]

† Grad. Sch. Systems and Information Engineering, University of Tsukuba,　Tsukuba, 305-8573, Japan
†† Research into Artifacts, Center for Engineering, University of Tokyo　Kashiwa, Chiba 277-8568, Japan
††† Navix Co., Ltd.　8-3-6 Nishi-Gotanda, Shinagawa-Ku Tokyo 141-0031, Japan
†††† Information Technology Center, University of Tokyo　7-3-1, Hongou, Bunkyo, Tokyo 113-0033, Japan
††††† National Institute of Informatics　2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

Abstract　This paper focuses on analyzing (Japanese) splogs based on various characteristics of keywords contained in them. We estimate the behavior of spammers when creating splogs from other sources by analyzing the characteristics of keywords contained in splogs. Since splogs often cause noises in word occurrence statistics in the blogosphere, we assume that we can efficiently collect splogs by sampling blog homepages containing keywords of a certain type on the date with its most frequent occurrence. We manually examine various features of collected blog homepages regarding whether their text content is excerpt from other sources or not, as well as whether they display affiliate advertisement or out-going links to affiliated sites. Among various informative results of our analysis, it is important to note that more than half of the collected splogs are created by a very small number of spammers.

Key words　blog, spam blog, affiliate

1. はじめに

ブログはパーソナル・ジャーナルとして意見情報が記されており，市場の動向を推測するための手掛かりや製品についての意見調査をする上で有益である．従来からあるインデクシングのみを行うサーチエンジンとは違い，ブログ特有の情報検索サービスが提供されている．具体的には，ブログ解析サービスとして，*Technorati*(注1)，*BlogPulse*(注2)，*kizasi.jp*(注3)，*blogWatcher* (注4) [1] などが存在する．多言語ブログサービスとしては，*Globe of Blogs*(注5) が言語横断ブログ記事検索機能を提供している．また *Best Blogs in Asia Directory* (注6) がアジア言語ブログの検索機能を提供している．*Blogwise*(注7)もまた多言語ブログ記事の分析を行っている．

一方で，ブログのウェブコンテンツの作成と配信は非常に簡単になっており，それはまたスプログの増加の要因ともなっている [2]〜[6]．スプログにおいては，機械的な文書作成や他サイトの引用という手段を用いて自動的に記事が生成され，広告主への誘導または対象サイトのページランクを増加する目的のリンクを有する. [4] は英語ブログにおいて，約 88%のブログサイトがスプログであり，それは全ブログポストの 75%を占めると報告した．この見積りに基づいて, [3], [7] に述べられているように，スプログは情報検索品質の低下やネットワークと格納資源の多大な浪費などといった問題を起こす要因となる．いくつかの既存研究 [4]〜[6] はスプログの重要な特性を報告している. [5] は TREC(注8)Blog06 データコレクションを用いて，スプログのピング時系列特性，入力度数/出力度数の分布特性，典型的な単語群を解析した. [4], [6] もまた *BlogPulse* データセットを用いたスプログ分析の結果を報告した. [4], [8], [9] は，スプログを機械的に特定し，排除することによって，スプログの引き起こす問題を解消するための研究である．

上記の既存研究とは異なり，本研究は日本語スプログの含むキーワードの様々な特性から，それに基づくスプログの分析に焦点を当てている．従来研究の多くが指摘している通り，スプログの本文は他者の作成したニュース記事，ブログ記事，広告ページ，その他のウェブ文書などから引用したものである．この知見より，スプログに含まれたキーワードの特性を分析することによって，他者の記事からスプログを作成するときのスパマーの嗜好を推測する [10]．また，スプログは，ブログにおける特定のキーワードの発生数の増加 (キーワードのバースト現象) に便乗をして，生成される傾向が強いので，キーワードがバーストする日に，そのキーワードを含むブログを収集することで，スプログを効率よく収集する方法を採用する．収集されたスプログは，内容文が他者の記事からの引用であるかそうではないか，またアフィリエイト広告やアフィリエイトサイトへのアウトリンクがあるかないかといったいくつかの特徴を人手で判別をした結果をデータセットとして蓄積している．データセットの分析を進めていく上で多くの有益な情報を得ることができたが，その中でも注目すべき重要な事実として，収集したスプログの半数以上が極少数のスパマーによって作成されていることがわかった．

(注1)：http://technorati.com/
(注2)：http://www.blogpulse.com/
(注3)：http://kizasi.jp/ (日本語のみ)
(注4)：http://blogwatcher.pi.titech.ac.jp/ (日本語のみ)
(注5)：http://www.globeofblogs.com/
(注6)：http://www.misohoni.com/bba/
(注7)：http://www.blogwise.com/
(注8)：http://trec.nist.gov/

2. スプログの作成手法

前章で述べたように，スプログの本文のほとんどはニュース記事，ブログ記事，広告ページ，その他のウェブ文書などの他者の記事から引用されたものである．しかしどのようなソースから作成されていても，スプログにはアフィリエイト指向があり，アフィリエイト広告またはアフィリエイトサイトへのアウトリンクを配置している．この指向のために，通常，スプログは，他者の記事の中で最新の内容を検索して，引用することによって，作成されている．このようなスプログの作成手順は大まかに以下の 2 つのケースに分けることができる．

i) **キーワード検索を用いずに**，最近のニュース記事またはブログ記事から引用

ii) **特定のキーワードを検索し**，それを含む他者の記事を引用

第 1 の手順で作成されたスプログ記事では，ごく最近のニュース記事やブログ記事に存在する最新の記事が盗用元である傾向がある．第 2 の手順で作成されたスプログ記事は，スパマーは，通常，ニュース記事やブログ記事から記事を検索するためのキーワードを吟味しており，高い効果のある adsense(注9)キーワードを選ぶ傾向がある．

3. スプログの素性

表 1 にまとめたように，本研究ではスプログ素性に対して考える要素は次の 3 つの観点によるタイプで大別する．

i) アフィリエイト性
ii) 本文の引用元
iii) 自動生成の手順

3.1 アフィリエイト性

[4], [6] で述べられているように，スプログは利益目的で広告を載せたり，不正にアフィリエイトサイトのランキングを向上させる目的によって作られている，偽のブログである．スプログはアフィリエイト広告に深く関連しているため，本研究において，アフィリエイト指向に関するスプログの素性を考えた．アフィリエイト性を表わす素性として，以下の 4 点を人手で判定した．

i) アフィリエイトサイトへのリンクの有無
ii) 広告の有無
iii) アダルトコンテンツの有無(注10)
iv) ポップアップ広告のキーワードへの自動埋め込みの

(注9)：http://google.com/adsense
(注10)：アフィリエイト広告の自体の主要なジャンルとして，健康食品，ダイエット商品，美容，金融などとともにアダルトコンテンツが挙げられる．アダル

表 1　スプログ素性及びそのスプログデータセット中での該当率

素性のタイプ	スプログ素性	説明	該当率 (%)
アフィリエイト性	アフィリエイトサイトへのリンク	ブログ記事内に，ブログホストが自動的に付与したもの以外にも，アフィリエイトサイトへのリンクが多く存在する.	80.5
	広告記事	ブログホストが自動的に配置したもの以外に，ブログ記事自体が広告文を多く含んでいる.	31.0
	アダルト記事	ブログ記事にアダルトコンテンツを含んでいる.	8.1
	ポップアップ広告	記事内のキーワードに自動的にポップアップ広告を付加する仕組みが存在する.	42.1
本文の引用元	ニュース記事の引用	本文の中に，ニュース記事からの自動・手動での引用がある.	14.3
	ブログ記事またはその他のウェブ文書の引用	本文の中に，他者のブログ記事，またはニュース記事や広告ページ以外のウェブ文書からの自動・手動での引用がある.	70.8
	広告ページからの引用	本文の中に，特定の広告ページからの自動・手動での引用がある.	27.1
	オリジナル文	スパマー自身がスプログの本文を書いている.	2.9
	無意味な単語列	主にワードサラダスパムテキスト [11] というものを指し，自動的に生成されている文章.	3.6
自動生成の手順	キーワード検索によらない引用	本文の中に，キーワード検索によらないで，他者の記事からの自動・手動での引用がある．通例，近い日付のニュース記事やブログ記事から引用をしている.	12.7
	日替わりのキーワード検索による引用	本文の中に，その日ごとのキーワードで検索をした，他者の記事からの自動・手動での引用がある.	49.5
	単一のキーワード検索による引用	本文全てが，単一のキーワードによる検索をした，他者の記事からの自動・手動での引用である.	36.9
	キーワード羅列 [6]	ブログ記事内に，SEO 目的の，キーワードの羅列を含む.	11.5
	自動生成文	主にワードサラダスパムテキスト [11] というもので，一見意味があるようで全く無意味な単語列を生成するものである．一部の文章は他者の記事からの引用である場合もある.	4.5

有無

3.2　本文の引用元

2. で説明したように，スプログの重要な特性として，スプログの本文のほとんどは，ニュース記事，ブログ記事，広告記事，その他のウェブ文書からの引用であるということである．スプログの生成の仕組みを推測するため，スプログの引用元を人手で判別し，本文の引用元を表わす素性として，以下の 5 点に分類をした.

i)　ニュース記事の引用
ii)　ブログ記事またはその他のウェブ文書の引用
iii)　広告ページからの引用
iv)　オリジナル文
v)　ワードサラダ [11] などの無意味な単語列

3.3　自動生成の手順

さらに，引用を行うためにウェブを検索する手順を推測し，自動生成の手順を表わす素性として，以下の 5 点を人手で判定した.

i)　キーワード検索をせずに，他者の記事の，最近のニュース記事やブログ記事などからの引用
ii)　その日ごとのキーワードで検索をした記事の引用
iii)　単一のキーワードで検索をしたブログ記事の引用
iv)　キーワード羅列 [6]
v)　ワードサラダ [11] を含む自動生成文

5.1 で述べたように，本研究の主旨はキーワード特性とスプログ素性の関係の分析にある．特に，本研究では，キーワードの時系列特性への影響と同様，特定のキーワードを用いてのスプログ生成手順の議論に焦点を当てている．自動生成の手順の素性において，以下の 2 点を区別する.

a)　**キーワード検索を用いずに**，最近のニュース記事またはブログ記事から引用
b)　**特定のキーワードを検索し**，それを含む他者の記事を引用

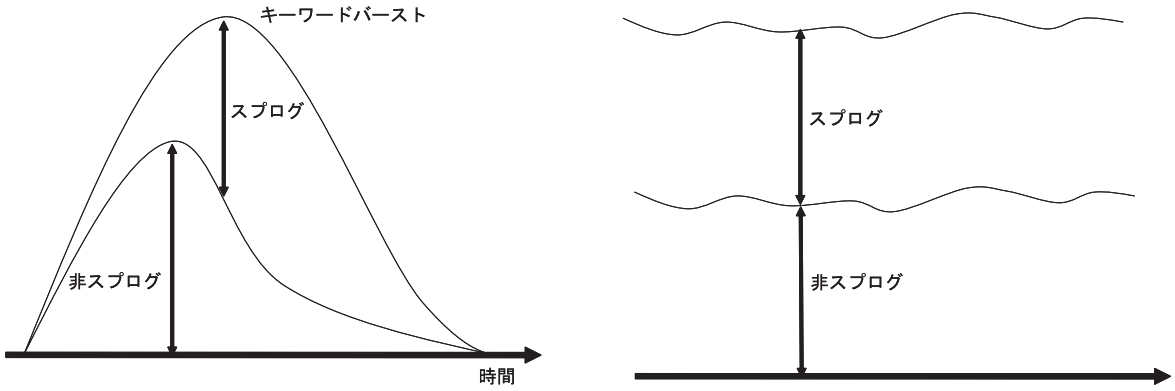
前者は上記の i) に，後者は ii)，iii) に対応をする．6. において，4.2 で述べたキーワード特性と，各キーワードで収集したスプログの素性の分布 (特に，上記 i) ∼ iii)) の関係を分析する.

4.　スプログとキーワードの特性

4.1　キーワードの時系列特性

この節では，ブログのキーワードバーストに便乗して，スプログが混入する現象について述べる．図 1 では時系列において，キーワード発生数の観測におけるスプログ混入の 2 つの典型的なケースを例示しており，(a) はバーストするキーワードの例，(b) はバーストの無いキーワードの例である．両者は各キーワードにおいてスプログと非スプログを分離する前の，混合した状態での出現頻度を表わしている．このままでスプログを検出・除去しなければ，非スプログだけの真のキーワードバース

トコンテンツを含むブログは他のジャンルのものよりも有害であるとみなし，一つの独立した素性として記録をしている.



(a) バーストするキーワード　　(b) バーストの無いキーワード

図 1　スプログ/非スプログにおける，キーワードの出現頻度の時系列特性

トを見積もることはできない．バーストするキーワードにおいては特に，スプログの本文がニュース記事やブログ記事などの他者の記事の引用であるということから，スプログのバーストが非スプログのバーストに遅れて発生すると推測される．

4.2　キーワードの特徴付けのためのキーワードマップ配置

この節では，図 2 に示すキーワード特徴付けのためのキーワードマップを紹介する．マップの縦軸は各キーワードの持つ情報価値を表わし，マップの横軸は各キーワードの情報有効時間を表わしている．情報価値の高いキーワードは，通常，社会・政治・経済分野のニュースとして報道されるものであり，情報価値の低いキーワードは，通常，娯楽や著名人に関する話題であるか，adsense キーワードとして高い効果を持つものである．また，情報有効時間の短いキーワードは，時期的なものや最新の出来事に関連したものであり，情報有効時間の長いキーワードは，歴史ある政党などの機関名や国名，または健康や美容などの恒久的な問題に関係するものである．

図 2 においては，マップ上で偏りなく分布するような，50 のキーワードを設定しているが，このようなキーワードの配置は完全に直観で決定している．これらの 50 のキーワードは，バーストする，しないといったキーワードの時系列特性において多様な特性を示す．また，これらのキーワードは，スプログ混入率においても多様な特性を取ることを狙って選ばれている．このようなマップに多様なキーワードを配置する狙いは，キーワード特性とスプログ混入率の関係を調べることにある．本研究では，これらの 50 のキーワードを用いて，スプログを収集する(注11)．

5.　キーワードの特性に基づくスプログ分析

5.1　キーワード特性に基づくスプログ解析の狙い

この節は，本論文の狙い，すなわち，キーワード特性に基づくスプログ分析の概要を述べる．具体的には，本研究ではブログを収集し，人手による分類判定を行った後，以下の 3 点についての分析の結果を報告する．

表 2　日本語ブログデータの概要 (2007 年 12 月 3 日 0:00 現在)

ブログ数	記事数	収集日数	最近の 1 日あたりの更新数
3,591,306	192,699,276	1,355	196,975

（1） スプログの作成手順の推測

（2） バーストするかしないかという，キーワードの時系列特性

（3） スプログの作成手順とキーワードの時系列特性の関係の分析．この分析は主に以下の要素を含む．

(3-a)　キーワード特性とスプログ混入率の関係．これはキーワード選択におけるスパマーの嗜好を明らかにするものである．

(3-b)　キーワード特性とスプログ生成手法の関係．

5.2　日本語ブログデータ

日本語ブログ収集にあたり，中国語，日本語，韓国語，英語のブログ記事を収集している，関心システム [13] を利用する．このシステムは各言語のブログサイトのリストを持っている．このリストより，ブログサイトの提供する RSS (注12) フィードファイルと Atom フィードファイルを取得し，形態素解析ツールを用いて，それらのフィードファイルからキーワードを抽出し，キーワードと記事をそれぞれのデータベースに蓄積している．このシステムでは，各言語でのブログ記事のインデクシングを行うため，いくつかの言語ツールを使用している．日本語においては，形態素解析ツールとして，Juman (注13) を用いている．関心システムは記事検索と分析をするための機能をユーザに提供している．

表 2 に，関心システムに蓄積されている日本語ブログデータの規模を示す．(2007 年 3 月現在) 2004 年からの集積により，360 万のブログサイト，1 億 9300 万記事の日本語ブログデータが蓄積されている．

5.3　分析の手順

この節では，5.1 で説明した主旨に従って，キーワード特性に基づいたスプログの収集と分析を行う手順を示す．まず，スプログ収集の戦略の概要を以下に示す．

(注11)：[12] では，スプログを含むウェブスパム全般について，スプログに現れやすいキーワードを収集した後，ウェブスパムを収集する手順が述べられている．

(注12)：略称の由来は，RDF Site Summary，Really Simple Syndication，Rich Site Summary など諸説ある．

(注13)：http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

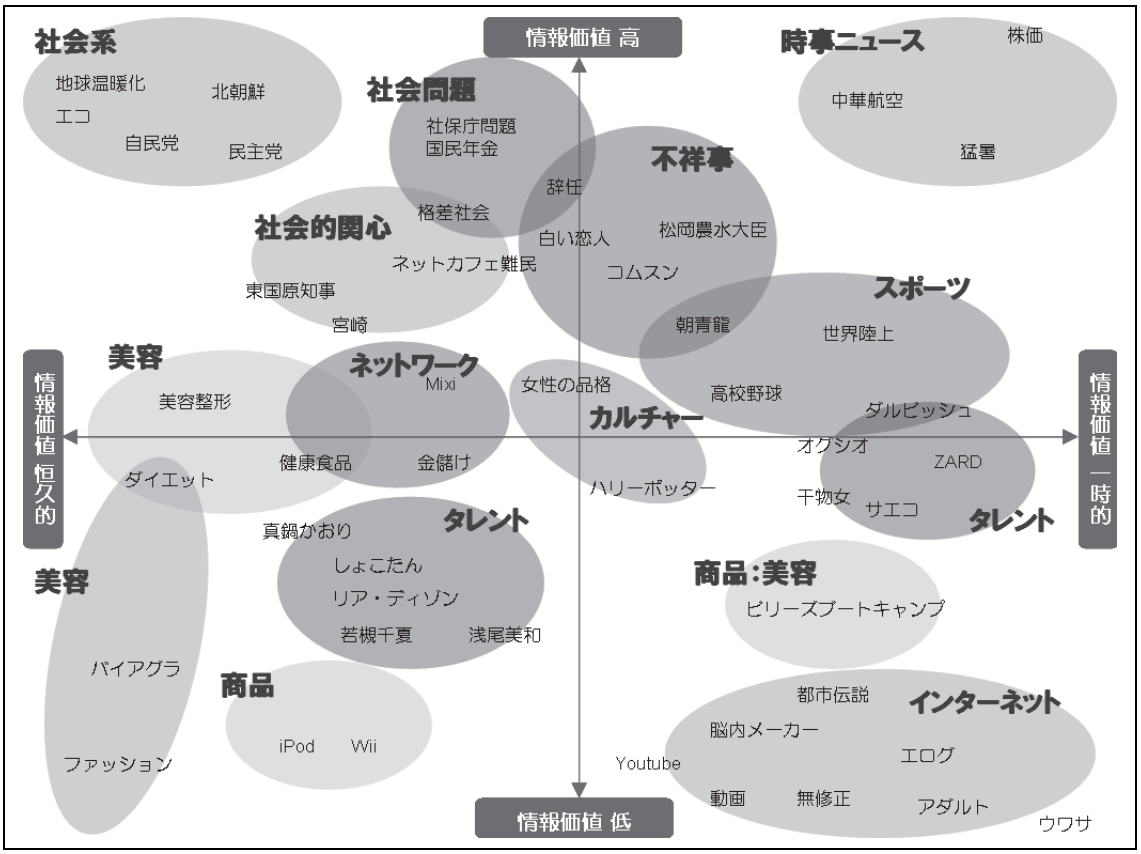


図 2　キーワード特性を表わすキーワードマップ

与えられたキーワードを含むブログサイトを収集し，各ブログサイトがスプログであるかないかを人手で判別し，さらに，3. で定義したスプログ素性を付与する．

ここで，バーストするキーワードにおいて，スプログの混入率はバースト日に増加する傾向がある．さらに，バーストの無いキーワードにおいても，スプログの混入率は，キーワードを含むブログサイト数が多い日に増加する傾向がある．また，多くのスプログは，1 日当たりの記事投稿数が非スプログよりも多い傾向がある．以上の観測に基づいて，効率的にスプログを収集するために，各キーワードにおいて，そのキーワードを含むブログサイト数が多い日を選び，1 日の記事投稿数の多いブログサイトを収集した．

以下に上記の手順の詳細を示す．

（1） 図 2 で示す 50 のキーワードにおいて，各キーワードを含むブログサイトの URL を，2007 年の最も投稿記事数が多い日で取得する．

（2） 取得した URL より，その日の投稿記事数で上位 50 件を選択し，さらにそれ以下から無作為に 60URL を選択した．合計 110 の URL の内，上位 50URL では，1 日の投稿数が 3 記事以上のものが集まる傾向があり，下位 60URL では，1 日の投稿数が 2 記事以下のものが集まる傾向がある．

（3） 取得した各 URL のブログサイトに対して，判定者がスプログ素性を付与する．

（4） 上の判定に基づいて，各 URL がスプログであるか非スプログであるかを以下のルールで決定する．

（a） その URL が以下の要件を満たすとき，それは**スプログである**と言える．

i. 「オリジナルの文章」が全く存在しない．

ii. 「オリジナルの文章」はあるが，「アフィリエイトサイトへのリンクがある」「広告記事がある」「アダルトコンテンツを含む」のいずれかを満たす．

（b） それ以外の場合，その URL は**スプログではない**．

（5） この結果，キーワード特性とスプログ素性の関係を分析する．

6. スプログデータセットの分析結果

4. 2 で述べた 50 キーワードの内，現在，図 4 に示す 22 キーワードの判定作業が終了しているため，この 22 キーワードでの初期評価を行った．

6. 1 ブログホスト会社の内訳

図 3 に示すように，全スプログの 88%は上位 3 件のホストに集中している．この内，上位 2 件のホストでは収集したブログ中のスプログ混入率は 50%前後であり，スプログ除去にかけているコストは他のホストよりも低いと思われる．また，ここで少数のスパマーが上位 3 件のスプログ混入率を示すホストにおいて，大量のスプログを生成していることが確認され，それらのホストにおけるスプログ混入率の増加に影響を与えていることが明らかになった．(注14)．

(注14)：[12] では，スプログを含むウェブスパム全般について，スパムが観測される Doorway ページ，Doorway ページからリンクされる広告サイト，イン

表 3　ホストごとのスプログ混入率

ホスト		S 社	C 社	J 社	A 社	L 社	G 社	Y 社	その他	計
ブログ数	スプログ	192	142	54	24	3	1	0	26	442
	非スプログ	203	115	169	355	128	130	207	396	1703
	計	395	257	223	379	131	131	207	422	2145
スプログ混入率 (%)		48.6	55.3	24.2	6.3	2.3	0.8	0.0	6.2	20.6

表 4　大量生成型スパマーの一覧

ID	件数	スプログ素性 (表 1 より)			キーワード
		アフィリエイト性	本文の引用元	自動生成の手順	
1	**115** (42.3%)	**サテライト、自動ポップアップ**	**ブログ**	**単一**	**ウワサ**、無修正、美容整形、**朝青龍**、サエコ、コムスン、ZARD、中華航空、北朝鮮、Wii、猛暑、干物女
2	**56** (20.6%)	**サテライト**	**ブログ**	**日替わり**	**エログ**
3	**30** (11.0%)	**サテライト**	**ニュース、(広告ページ)**	**キーワード無し**	**国民年金**、コムスン
4	**26** (9.6%)	**サテライト、アフィリエイトリンク、(自動ポップアップリンク)**	**ブログ、広告ページ**	**日替わり**	**国民年金**
5	**20** (7.4%)	**サテライト、(アフィリエイトリンク)**	**広告ページ**	**日替わり、羅列**	**健康食品**
6	10 (3.7%)	サテライト、アダルト、自動ポップアップ	ニュース、ブログ	キーワード無し	**エログ、朝青龍**
7	7 (2.5%)	サテライト、(アフィリエイトリンク)	広告ページ	日替わり、羅列	**健康食品**
8	4 (1.5%)	サテライト、アダルト、自動ポップアップ	ニュース、ブログ	キーワード無し	**エログ**
9	2 (0.7%)	サテライト	広告ページ	単一	バイアグラ
10	2 (0.7%)	サテライト	広告ページ	日替わり	美容整形
計	272	-	-	-	-

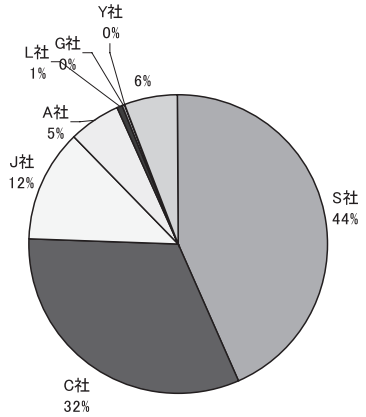


図 3　スプログデータセット中のブログホストの分布

6.2　キーワード素性とスプログ素性の関係

次に，22 キーワードのスプログ混入率を表 5 に降順に並べた．表 5 より，22 キーワードをスプログ混入率の大きさによって，30%超，30 ～ 10%，10%未満の 3 つのグループに分けた．なお時系列特性を示すため，バーストの無いキーワードには下線を引いた．さらに，データセット中の全てのスプログについての，各スプログ素性の頻度を計上し，全スプログ中の該当率を表 1 の最右列に記した．このスプログ素性の分析に基づいて，スプログ素性と，スプログ混入率 10%超のキーワード特性の関係を調べた．

そして，構造の酷似するスプログ同士が，同一のスパマーによって作られたものであるかを判定し，それらの同一スパマーによるスプログを他のスプログと区別し，同一スパマーごとに分類した．この分類により，データセット中の全スプログ 442 件の中で，2 件以上のスプログを作成したスパマー 10 人を同定できた．また，全スプログ 442 件の中で，272(61.5%) がこれらの 10 人のスパマーによって作られたものであると分類できた．本研究では，特に，これらの 10 人のスパマーを「大量生成型」スパマーと呼び，その他のスパマーを「単発」スパマーと呼ぶ．これらの「大量生成型」スパマーの特徴の概要を表 4 に示す．この「大量生成型」スパマーの判定に基づいて，各キーワードによって収集したスプログの中で,「大量生成型」スパマーの手によるスプログの占める割合を算出した．また，データセット全体から,「大量生成型」スパマーの手によるスプログを除外し,「単発」スパマーと非スパムのブログのみによるスプログ混入率を算出し直した．それらの結果を表 5 に示す．

この分析の結果の概要として，2 次元マップ上に表わしたものを図 4 に示す．

(1) スプログ混入率が 30%超のキーワードの 5 件中 4 件のスプログの大部分は「大量生成型」スパマーの手によるものであ

ターネット・サービス・プロバイダー，広告仲介業者，広告主の五階層に渡って，ウェブスパムとの関係が深いドメインの統計分析結果を示している．

表 5　キーワード別スプログ混入率 (下線：バースト無し，**太字**：「大量生成型」スプログ率 50%超)

キーワード	スプログ率 (%)	スプログ中の「大量生成型」スプログ率 (%)	「大量生成型」スパマー ID	「大量生成型」スプログ除去後のスプログ率 (%)
エログ	89.2	**92.4**	2, 6, 8	38.5
ウワサ バースト無し	88.1	**94.8**	1	27.8
国民年金	58.1	**90.2**	3, 4	12.0
無修正 バースト無し	40.9	18.5	1	36.1
健康食品 バースト無し	37.4	**58.7**	5, 7	19.8
美容整形 バースト無し	24.4	14.3	1, 10	21.7
バイアグラ バースト無し	22.5	11.1	9	20.5
ダルビッシュ	22.1	0.0	-	22.1
動画 バースト無し	19.1	0.0	-	19.1
朝青龍	15.2	**80.0**	1, 6	3.4
ビリーズブートキャンプ	15.1	0.0	-	15.1
サエコ	14.3	14.3	1	12.2
コムスン	6.9	71.4	1, 3	2.1
ZARD	4.7	20.0	1	3.8
中華航空	4.7	20.0	1	3.8
北朝鮮	2.9	100.0	1	0.0
Wii	2.8	66.7	1	1.0
猛暑	2.8	33.3	1	1.9
女性の品格	2.0	0.0	-	2.0
干物女	1.8	50.0	1	0.9
参議院選挙	0.0	0.0	-	0.0
民主党	0.0	0.0	-	0.0
計	20.5	61.5	1 - 10	9.0

る．このキーワードのスプログは本研究で収集したスプログの 6 割超を占める．スパマーがいつスプログを生成し，どのキーワードを選ぶかによって，ここに現れるキーワードやスプログの素性は大きく影響を受けるものと思われる．

(2) 図 4 より，スプログ混入率が 10%未満のキーワードはほとんどマップ上半分に配置されているとわかる．これより，スプログには情報価値の高いキーワードより情報価値の低いキーワードが含まれる傾向があると言える．例外的にスプログ混入率が 30%超で上半分に配置するキーワードである国民年金・朝青龍は,「大量生成型」スパマーの影響を大きく受けている．これは「大量生成型」スパマーがニュース記事の盗用という仕組みを選んでスプログを生成した時期に，偶然，国民年金・朝青龍に関連する報道が多かったため，結果的にこれを含むスプログが多数生成されたことによる．

(3) ウワサ，エログ，健康食品の 3 件のキーワードは，(2) とは別の「大量生成型」スパマーと関連している．これらのキーワードにおけるスプログでは，他ブログまたは広告文の引用が多く，ニュースの引用は少ない．

(4) スプログ混入率 10%超でバーストするの 6 キーワードの内，エログだけは例外であり，収集の期間において，単にブログホスト側がこのキーワードを含むスプログの削除を行ったため，バーストを起こしているように見えただけである．

(5) スプログ混入率 30%超の 5 キーワードの内，無修正以外の 4 キーワードは「大量生成型」スパマーの手によるスプログの占める割合が高く，強い影響を受けている．

(6) 「大量生成型」スパマーの影響を除外して再計算したスプログ率がなお高いキーワードの多くは，バースト性の低い恒常的なキーワードである．

7. ま と め

本研究ではキーワードのバースト特性に基づいて日本語スプログの収集・分析を行った．収集したスプログデータセットにおいて，半数以上のスプログは，少数のスパマーが生成している事が判明した．今後の展開として, [4]〜[6] で研究された，スプログ中の特徴語，入出次数分布，ピング時系列などの特徴を含めて更なる分析を進める．次に，データセットに蓄積されたスプログ判別例を基に，既存のスプログ検出技術 [7], [9] を適用して，高精度のスプログ判別器を開発し，さらなるデータセットの拡張に役立てる．

文　献

[1] T. Nanno, T. Fujiki, Y. Suzuki, and M. Okumura. Automatically collecting, monitoring, and mining Japanese weblogs. In *WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pp. 320–321. ACM Press, 2004.

[2] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *AIRWeb '05: Proceedings of the 1st International Workshop on Adversarial Information Retrieval on the Web*, pp. 39–47, 2005.

[3] *Wikipedia, Spam blog.* `http://en.wikipedia.org/wiki/Spam_blog`.

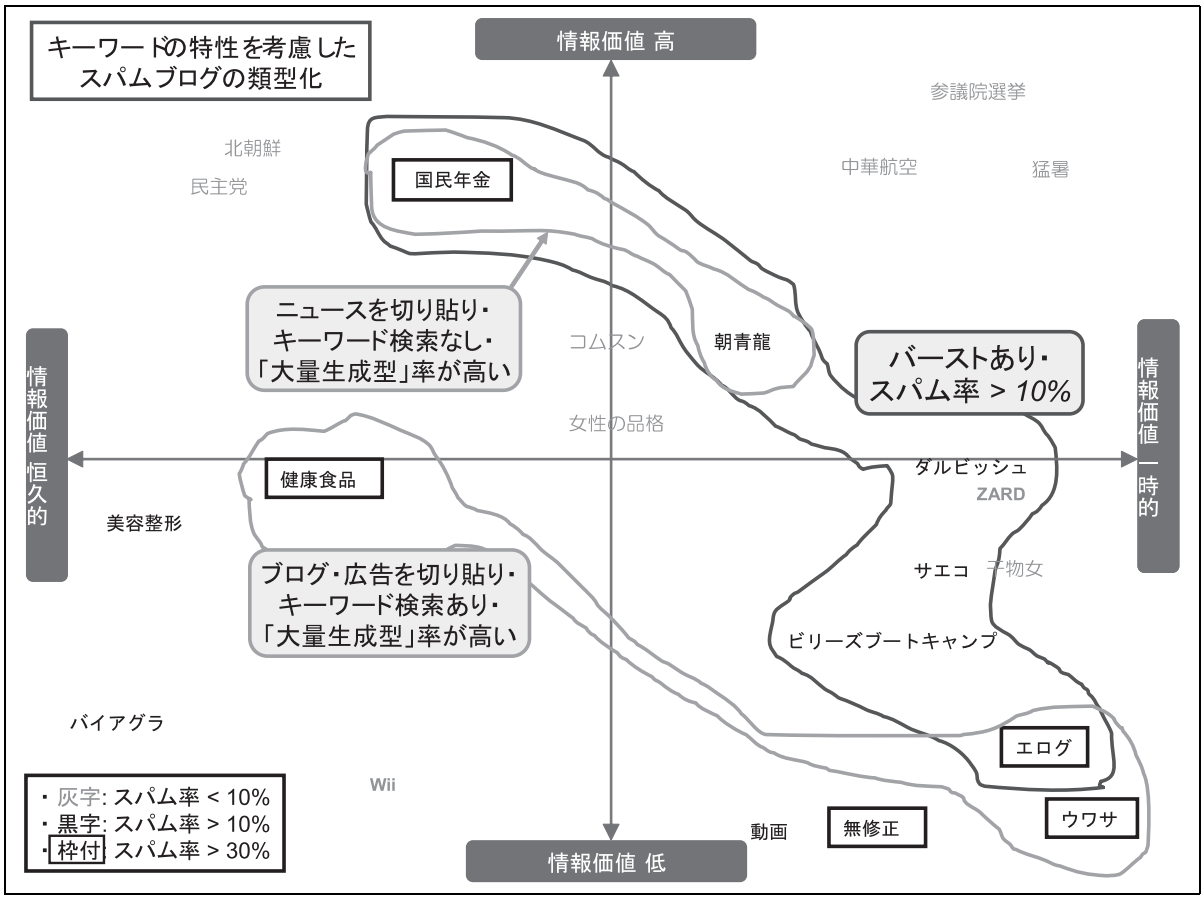


図 4　キーワードマップ上で見るスプログ分析の結果

[4] P. Kolari, A. Joshi, and T. Finin. Characterizing the splogosphere. In *Proceedings of WWW 2006 3rd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2006.

[5] C. Macdonald and I. Ounis. The TREC Blogs06 collection : Creating and analysing a blog test collection. Technical Report TR-2006-224, University of Glasgow, Department of Computing Science, 2006.

[6] P. Kolari, T. Finin, and A. Joshi. Spam in blogs and social media. In *Tutorial at ICWSM*, 2007.

[7] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. L. Tseng. Splog detection using self-similarity analysis on blog temporal dynamics. In *AIRWeb '07: Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web*, pp. 1–8, 2007.

[8] 石田和成. スパムブログの定量的調査と分離の試み. データベースと Web 情報システムに関するシンポジウム (DBWeb2007) 論文集. 情報処理学会, 2007.

[9] P. Kolari, T. Finin, and A. Joshi. SVMs for the Blogosphere: Blog identification and Splog detection. In *Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, pp. 92–99, 2006.

[10] Y. Sato, T. Utsuro, T. Fukuhara, Y. Kawada, Y. Murakami, H. Nakagawa, and N. Kando. Collecting and analyzing Japanese splogs based on characteristics of keywords. In *Proceedings of ICWSM*, 2008.

[11] *Wikipedia, Word salad (computer science).* `http://en.wikipedia.org/wiki/Word_salad_%28computer_science%29`.

[12] Y.M. Wang, M. Ma, Y. Niu, and H. Chen. Spam double-funnel: Connecting web spammers with advertisers,. In *Proceedings of the 16th WWW Conference*, pp. 291–300, 2007.

[13] 福原知宏, 宇津呂武仁, 中川裕志. 複数言語間の語彙出現傾向比較による言語横断型ウェブログ関心解析システムの開発. 言語処理学会第 13 回年次大会「大規模 Web 研究基盤上での自然言語処理・情報検索研究」 ワークショップ論文集, pp. 40–43, 2007.