3E2-1

キーワードの特性を利用したスパムブログの収集と分析

Collecting/Analyzing Splogs based on Characteristics of Keywords

佐藤 有記[*1] Yuuki Sato　宇津呂 武仁[*1] Takehito Utsuro　福原 知宏[*2] Tomohiro Fukuhara　河田 容英[*3] Yasuhide Kawada　村上 嘉陽[*3] Yoshiaki Murakami
中川 裕志[*4] Hiroshi Nakagawa　神門 典子[*5] Noriko Kando

[*1]筑波大学大学院システム情報工学研究科 Grad. Sch. Sys & Info Engineering, Univ. Tsukuba
[*2]東京大学 人工物工学研究センター Research into Artifacts, Center for Engineering, Univ. Tokyo
[*3](株) ナビックス Navix Co., Ltd.
[*4]東京大学 情報基盤センター Information Technology Center, Univ. Tokyo
[*5]国立情報学研究所 National Institute of Informatics

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.

1. はじめに

ブログや掲示板などのウェブ上に大量に存在するテキストデータからの意見抽出研究の上で，スプログはノイズとして障害を引き起こす存在であり，これらを機械的に特定し，排除するべき対象である．[石田 07, Kolari06] はそのためのスプログ自動検出の研究である．ここで，スプログはアフィリエイトの目的のため，他者の記事を盗用しながら機械的に生成され，大量複製されるブログであり，ユーザを誘引するためのキーワードを設定しているものと推測される．

本研究はスパムブログデータセットを効率的に集めることを目的として，キーワードのバースト特性を利用して，スパムブログを収集し，データセットを作成する．キーワード特性とスプログの素性との間の関係について分析し，スパマーの嗜好の分析を行う [佐藤 08b, Sato08a]．収集した一部データセットの解析から，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．

2. スプログの素性

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[*1] キーワードを選ぶ傾向がある．

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

それぞれの素性を人手で判定した．その判定結果を蓄積したデータセットより，5. において，3. 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．

3. キーワードの特性

この章では，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関心の公私性を表わし，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．公的関心性が高いキーワードは，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，私的関心性が高いキーワードは，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，情報有効時間とは，時系列においてのキーワード発生数のバースト性を表わす軸であり，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，キーワードの出現頻度がバーストするものである．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．

連絡先: 佐藤有記，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，ysato@nlp.iit.tsukuba.ac.jp

*1 http://google.com/adsense

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	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

本研究で扱うキーワードは，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．この 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．

4. キーワードの特性に基づくスプログ分析

4.1 キーワード特性に基づくスプログ解析の狙い

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．

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．

4.2 分析の手順

スプログ収集の戦略の概要を以下に示す．

> 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，2. 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．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 [福原 07] を利用する．

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．

5. スプログデータセットの分析結果

3. で述べた 50 キーワードの内，現在，図 2 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．

5.1 ブログホスト会社の内訳

図 1 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．

表 2: 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 3: 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∼10	15	(5.5%)	—	—	—	**エログ，健康食品**，バイアグラ，美容整形
計	272		—	—	—	—

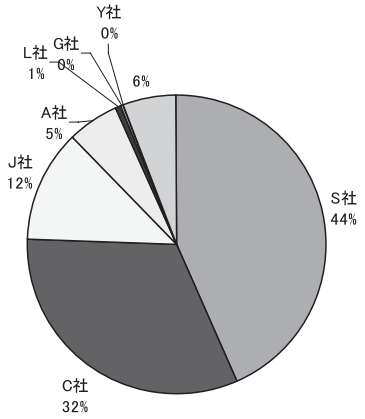

図 1: スプログデータセット中のブログホストの分布

5.2 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 4 に降順に並べ，その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 3 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 4 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 2 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 2 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 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．
5. 「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．

表 4: 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，ZARD，中華航空，北朝鮮，Wii，猛暑，女性の品格，千物女，参議院選挙，民主党はスプログ混入率 10%未満				
計	20.5	61.5	1 - 10	9.0

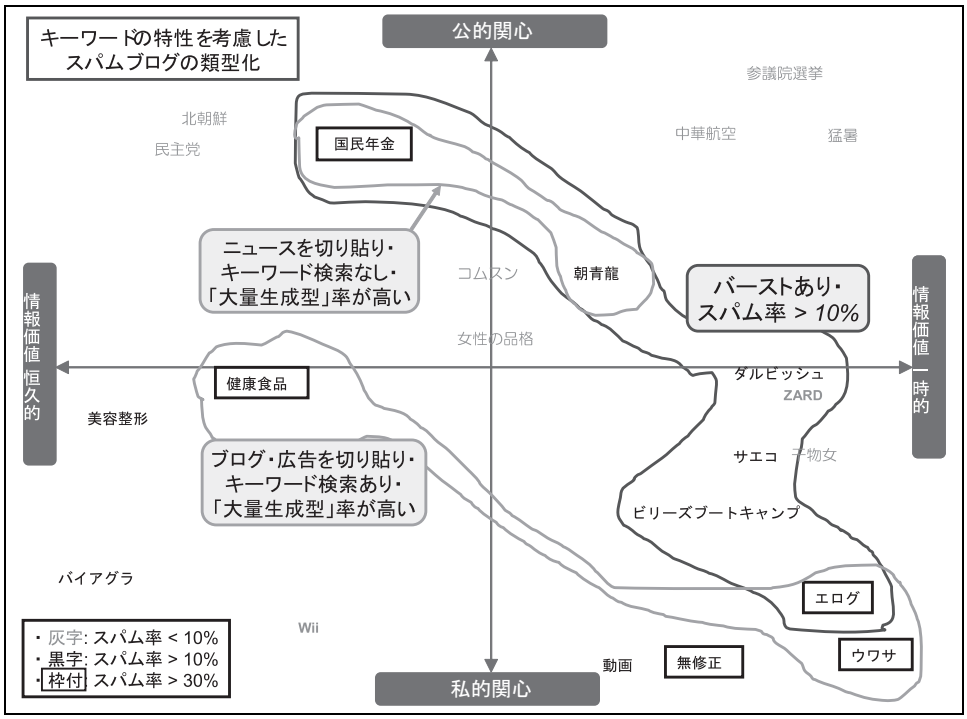

図 2: キーワードマップ上で見るスプログ分析の結果

6. 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，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 [Kolari06] 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．

参考文献

[福原 07] 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).

[石田 07] 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).

[Kolari06] Kolari, P., Finin, T. and Joshi, A.: SVMs for the Blogosphere: Blog Identification and Splog Detection, *Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, pp. 92–99 (2006).

[Sato08a] Sato, Y., Utsuro, T., Fukuhara, T., Kawada, Y., Murakami, Y., Nakagawa, H. and Kando, N.: Analysing Features of Japanese Splogs and Characteristics of Keywords, *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web* (2008).

[佐藤 08b] 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：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, 電子情報通信学会第 19 回データ工学ワークショップ, 第 6 回日本データベース学会年次大会 (DEWS2008) 論文集 (2008).