キーワードのバースト特性を利用したスパムブログデータセットの作成と分析*

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†，東京大学 人工物工学研究センター‡
(株) ナビックス§，東京大学 情報基盤センター¶，国立情報学研究所‖

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．[2–4] 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 [5]．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 データセット作成手法

2.1 キーワードの素性と設定

キーワードの持つ素性として情報価値・情報有効時間を設定した 2 次元マップ (図 2) 上の配置を考え，特性の分布に偏りの無いよう 50 のサンプルキーワードを設定した．

2.2 キーワードの時系列特性

ブログにおいて，キーワードがバーストする際，そこにはスプログによるノイズが含まれていることが多い．よってキーワードを含むポストの頻度の時系列的変化を福原らの関心システム [1] にて観測する．このシステム中の日本語ブログデータセットより，2007 年において，ブログ中にキーワードを含む記事数が最大となるバースト日にポストしているブログページから 110 件をサンプリングし取得した．

2.3 スプログの素性

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 [4]．スプログの作成手法の多くは他者の作成した文章を盗用元として機械的に複製をし，スパマー自身はコンテンツ制作にほとんど労力をかけない．しかし機械的複製をする上でも，よりアフィリエイト効果を高めるべく，特定のユーザを誘引する目的でそのユーザを誘引できるキーワードをスプログに設定しているものが存在する．ニュース記事を盗用することで，最新のトピックに興味を持つユーザを無差別的に誘引するものや，特定のキーワード，特に高い効果の adsense [1] キーワードを含むポストを収集して，特定のユーザを誘引し，アフィリエイト効果を期待しているものもある．本研究では取得したスプログのデータに対して以上の観点からの詳細な分類を行う [5]．

i) アフィリエイト広告やアフィリエイトサイトへのリンクの有無
ii) スプログ本文の盗用元
iii) スプログ本文の自動生成の手順

3 スプログデータセットの分析結果

現在 50 キーワード中 22 キーワードの判定作業が終了しているため，22 キーワードでの初期評価を行った．

3.1 ブログホスト会社の内訳

図 1 の示すように，全スプログの 85%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%以上であり，スプ

*Collecting and Analyzing Japanese Splog Data Set based on Burst of Keywords

†Yuuki Sato, Takehito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba,
‡Tomohiro Fukuhara, Research into Artifacts, Center for Engineering, University of Tokyo,
§Yasuhide Kawada, Yoshiaki Murakami, Navix Co., Ltd.
¶Hiroshi Nakagawa, Information Technology Center, University of Tokyo,
‖Noriko Kando, National Institute of Informatics

[1] http://google.com/adsense

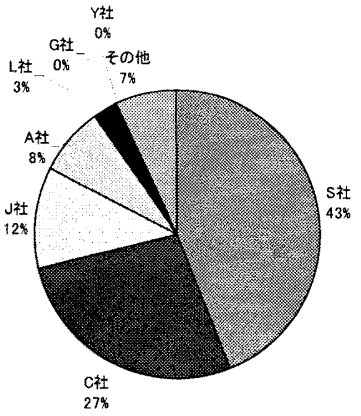

図 1: スプログデータセット中のブログホストの分布

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，ここで特定少数のスパマーが大量のスプログを生成していることが確認された．

3.2 キーワード素性とスプログ素性の関係

22 キーワードのスプログ混入率を表 1 に，2 次元マップ上の配置を図 2 に示す．

(1) 40%超のキーワードの 5 件中 4 件のスプログのほとんどは特定少数の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半数を占める．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，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 40%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 5 人以下の特定少数のスパマーの影響を大きく受けている．これは特定少数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 件では他ブログまたは広告文の引用が多く，ニュース引用は少ない．

4 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，特定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] 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 [3] 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．

表 1: キーワード別スプログ混入率

キーワード	スプログ混入率 (%)
ウワサ バースト無し	**88.2**
エログ	79.1
国民年金	58.9
無修正 バースト無し	**57.0**
健康食品 バースト無し	**43.0**
動画 バースト無し	**27.5**
バイアグラ バースト無し	**27.8**
ビリーズブートキャンプ	23.4
ダルビッシュ	22.9
美容整形 バースト無し	**20.9**
朝青龍	17.5
サエコ	16.4
中華航空，コムスン，ZARD，女性の品格，猛暑，Wii，北朝鮮，干物女，参議院選挙，民主党はスプログ混入率 10%未満	
計	23.5

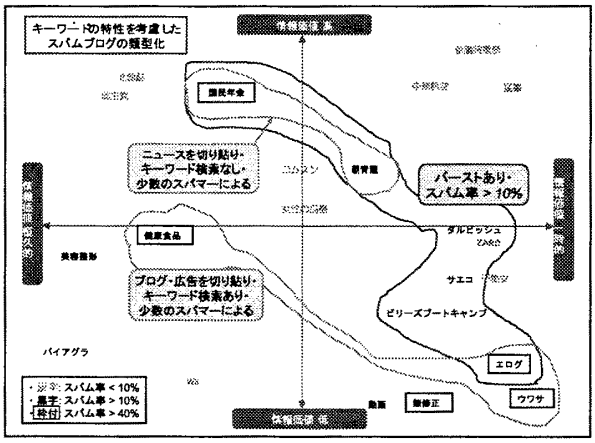

図 2: キーワードマップ上で見るスプログ分析の結果

参考文献

[1] 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.

[2] 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.

[3] 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. AAAI Press, 2006.

[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] 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.