

# ユーザ適応型 Splog フィルタリングのためのユーザ固有 Splog 空間の分析

芳中 隆幸<sup>†</sup> 福原 知宏<sup>††</sup> 増田 英孝<sup>†</sup> 中川 裕志<sup>†††</sup>

<sup>†</sup> 東京電機大学未来科学部情報メディア学科 〒101-8457 東京都千代田区神田錦町 2-2

<sup>††</sup> 東京大学人工物工学研究センター 〒277-0882 千葉県柏市柏の葉 5-1-5

<sup>†††</sup> 東京大学情報基盤センター 〒113-0033 東京都文京区本郷 7-5-1

E-mail: †yoshinaka@cdl.im.dendai.ac.jp, ††fukuhara@race.u-tokyo.ac.jp, †††masuda@im.dendai.ac.jp,  
††††nakagawa@dl.itc.u-tokyo.ac.jp

あらまし 今日のスパムブログ (Splog) は増加の一途を辿っており, Web やブログ空間におけるノイズとなっている. 本研究では, 効果的な Splog フィルタリングの手法としてユーザ適応型 Splog フィルタリングの構築を目指し, その必要性を検証するため実験システムを用いた評価を行った. その結果, ユーザ固有の Splog 空間の存在がわかった.

キーワード スパム, ブログ, Splog フィルタリング, ユーザ適応

## Analysis of Individual Splogosphere for User Adaptable Splog Filtering

Takayuki YOSHINAKA<sup>†</sup>, Tomohiro FUKUHARA<sup>††</sup>, Hidetaka MASUDA<sup>†</sup>, and Hiroshi  
NAKAGAWA<sup>†††</sup>

<sup>†</sup> Tokyo Denki University, 2-2 Kanada-Nishiki-cho, Chiyoda-ku, Tokyo 101-8457

<sup>††</sup> The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-0882

<sup>†††</sup> The University of Tokyo, 7-5-1 Hongo, Bunkyo-ku, Tokyo 113-0033

E-mail: †yoshinaka@cdl.im.dendai.ac.jp, ††fukuhara@race.u-tokyo.ac.jp, †††masuda@im.dendai.ac.jp,  
††††nakagawa@dl.itc.u-tokyo.ac.jp

**Abstract** Today, spam blog sites (Splogs) increase on the Web and the blogospheres so these splogs become the noise. We describe analysis results of individual splogospheres based on user studies, towards the development of user adaptable splog filters as effective approach. We had a survey using a splog check system called SplogExplorer. Analysis results of individual splogospheres are described.

**Key words** spam blog (Splogs), web spam filtering, user adaptation

### 1. はじめに

今日, 情報発信の手段としてブログが普及している. その一方で, 商品の宣伝を目的としたスパムブログ (Splog(スブログ)) が増加し, 様々な問題を引き起こしている. また, Splog の種類は多様であり, 定義 [1] は存在するが, その認知度の低さから, ユーザごとにおける Splog への判断は異なる傾向にある [2]. そこで, 本研究ではユーザ適応型 Splog フィルタリングを提案し, システムを用いた評価実験を行った.

以下, 本論文の構成は次の通りである. 2. では, Splog フィルタリングの現状と Splog 定義について先行研究からレビューを行う. 3. では, 評価実験における実験結果とその分析, 考察を行う. 4. では, 本論文のまとめと今後の課題について述べる.

### 2. Splog フィルタリングと Splog 定義の現状

#### 2.1 Splog フィルタリングの現状

Kolari は英語圏の Splog 空間に対して調査を行い, Splog 検知手法として SVM(Support Vector Machine) を取り入れ F 値約 70% の Splog 検知に成功している [1]. 日本語圏における Splog 研究としては石田の研究 [3] があり, Splog フィルタリングのための特徴として「リンク」に着目した研究を行ない 90% の Splog 検知に成功している.

これら先行研究では, 高い Splog 検出率を示しているが, 提案している Splog フィルタリングはユーザ間で共通に存在する 1 つの Splog フィルタリングであり, 個々のユーザに対応できるような柔軟なフィルタリングではない. そこで, 我々はユーザ適応型フィルタリングの提案を行ない, ユーザごとに興味を

反映できる柔軟なフィルタの開発を目指す。

## 2.2 Splog 定義

日本語圏の Splog 空間における Splog 定義では、主に 3 タイプに分類される Splog 定義 [2] がある。本論文では上記 3 タイプを 4 タイプに拡張し Splog 定義として位置づける。以下にその 4 タイプの定義を示す。

- (1) アフィリエイト型 Splog
- (2) コピー&ペースト (コピペ) 型 Splog
- (3) ワードサラダ型 Splog
- (4) アダルト型 Splog

また、本論文では上記 4 タイプに該当しないものを非 Splog とする。

## 3. 評価実験

本評価実験における目的はユーザ固有 Splog 空間の存在を示すことである。以下、評価実験における結果と考察について述べる。

### 3.1 評価実験環境および実験方法

- 実験方法 … テスト記事 50 件に対して被験者 12 名が 3.2 節に示す SC システムを実際に使用し Splog か非 Splog を判定する。

- テスト記事 (50 件) … 本 Splog 定義を元に筆者が Splog と非 Splog を含むブログ記事 50 件をを選定した。内訳は Splog: 非 Splog=35:15 とし、これらをベースライン (BL) として設定する。

- 被験者 … 被験者は工学系大学生 12 名で男女比は 11:1、年齢は全員 20 代である。

- 事前説明 … 被験者には本論文で採用している Splog 定義について事前説明を行う。これにより「Splog 定義に準拠した判定」というタスクを被験者に設ける。

### 3.2 評価実験システム:SplogChecker

本研究では、ユーザ適応型 Splog フィルタリングの実現のため、支援ツール SplogChecker(SC)を開発した [2]。SC は、Splog データセット作成支援ツールとして開発され、ユーザ固有で Splog 空間を管理することができる。

### 3.3 システム利用による評価実験

本実験結果から被験者ごとに Splog 空間が存在していることがわかった。テスト記事 50 件における被験者ごとの Splog 判定記事数の分布を図 1 に示す。縦軸にはテスト記事数 (50) をとり、横軸には被験者数 (A~L) をとっている。図 1 内の数値はテスト記事中における被験者の Splog 判定記事数を示している。また、図 1 は左から Splog 判定記事数でソートしているため右に行くほど Splog 判定記事数が少ない被験者となっている。図 1 中、各グラフの中央部分 (BS-nS と Bns-S の合成部分) が被験者判定と BL 判定との不整合部分を表しており<sup>(注1)</sup>この部分が広いほど BL と比較したときの被験者の判定精度が低いことを示している。

図 1 において、テスト記事 50 件のうち最も多い Splog 判定記事数だったのは E の被験者で 42 件もの Splog があると判定

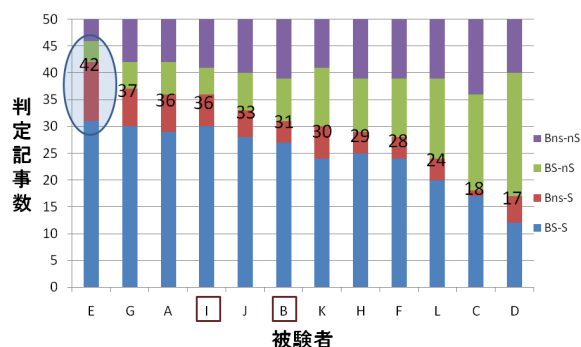


図 1 被験者ごとによる Splog 判定記事数の分布 (図中の数字は 50 件のうち Splog と判定した数) (図中の円で囲まれたのが BL との不整合部分) ・ 凡例説明

Bns-nS … BL 判定, 被験者判定ともに非 Splog  
 BS-nS … BL 判定は Splog, 被験者判定は非 Splog  
 Bns-S … BL 判定は非 Splog, 被験者判定は Splog  
 BS-S … BL 判定, 被験者判定ともに Splog

している。対して最も少ない Splog 判定記事数だったのは D の被験者で 17 件しか Splog と判定していない。BL を正解データとして見た場合の各被験者の判定精度では、B と I の被験者が 0.732 と最も高かった。また、被験者全員が Splog, 非 Splog と判定した記事がそれぞれ 7 件、1 件存在し、前者はアフィリエイト型とアダルト型 Splog 記事であり、後者は日記型の非 Splog 記事であった。一方、被験者で判定が大きく割れた記事 (Splog:非 Splog=6:6) も 3 件存在し、それらはアフィリエイト型の Splog 記事であった。以上のことから、Splog 空間にはユーザに依存しない共通 Splog 空間が存在すると同時に、アフィリエイト型のようなブログ記事でもそれを有益な情報とするユーザが存在することから、ユーザ固有の Splog 空間の存在がわかる。

## 4. おわりに

本研究では、ユーザ適応型 Splog フィルタリングの提案とその必要性の検証として、システムを利用した評価実験を行った。結果として、ユーザ固有の Splog 空間が存在していることがわかり、本論文で提案するユーザ適応型 Splog フィルタリングの必要性を示すことができた。今後は提案しているユーザ適応型 Splog フィルタリングの本実装、評価を行い、フィルタの実装にはユーザの Splog 判定を学習できるような機能を備えた実装を行いたいと考えている。

## 文 献

- [1] Pranam Kolari, Tim Finin, Akshay Java, and Anupam Joshi. SVMs for the Blogosphere: Blog Identification and Splog Detection. *In Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, pages 92–99, March 2006.
- [2] 芳中 隆幸, 福原 知宏, 増田 英孝, and 中川裕志. ユーザ適応型 splog フィルタリングに向けた splog 空間調査ツールの開発と評価実験. 電子情報通信学会第 12 回 Web インテリジェンスとインタラクション研究会, pages 59–64, Jul 2008. WI2-2008-37.
- [3] 石田 和成. スパムブログを除いたソーシャルメディアにおけるマスメディアの影響分析. FIT2007 第 7 回情報科学技術フォーラム, Sep 2008. D-038.

(注1): 例えば E の被験者では図の円で囲まれている部分