

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

芳中隆幸 増田英孝[†] 福原 知宏^{††} 中川 裕志^{†††}

[†] 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

^{††} 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

^{†††} 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, masuda@im.dendai.ac.jp, ††fukuhara@race.u-tokyo.ac.jp,
†††nakagawa@dl.itc.u-tokyo.ac.jp

あらまし 今日ブログツールやブログサービスの普及によりブログは情報発信の手段として広く一般的に使用されるようになった。しかしその一方で、ブログサイトの中にはスパムブログ (Splog) が増加し、検索におけるノイズとなっている。そこで本研究では、Splog 空間の調査分析ツールとして SplogExplorer の開発を行い、Splog 空間の調査を行う他、ユーザ適応型 Splog フィルタリングの概念を提案し、実装に向けた予備実験結果について報告する。

キーワード スパムブログ (Splog), Splog フィルタリング, ユーザ適応

Takayuki YOSHINAKA, HIDETAKA MASUDA[†], Tomohiro FUKUHARA^{††}, and Hiroshi
NAKAGAWA^{†††}

[†] Tokyo Denki University, 2-2 Kanada-Nishiki-cho, Chiyoda-ku, Tokyo 101-8457

^{††} The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-0882

^{†††} The University of Tokyo, 7-5-1 Hongo, Bunkyo-ku, Tokyo 113-0033

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

Abstract Today, blog tools and services become popular. The blog came to be used widely and generally as the information sending. However, spam blog (Splog) articles increase in blogosphere, various problems are caused by splogs. Splogs hinder users for searching the Web because Splogs are noises in the search result. In this paper, we suggest user adaptable of Splog filtering and report on the preliminary experiment result. The evaluation of splogs is different in each user, and user really needs the user adaptable filtering because it exists as a peculiar space to the user. In this research, to collect the needed data in the user adaptable filtering, the tool : SplogExplorer was developed.

Key words SpamBlog, SplogFiltering, User Adaptable

1. はじめに

今日、ブログツールやブログサービスの普及に伴い、多くの人がブログサイトを開設し、情報発信できるようになった。一方、ブログサイトの中には、価値の低いブログサイトやスパムブログ (Splog(スブログ)) が増加し、検索エンジンにおける不当な順位操作や検索結果における精度低下の原因となっている。

本論文では、Splog とは何かを明確にするため独自の Splog 定義付けを行う。Kolari らは英語圏の Splog 研究において

Wikipedia^(注1)などから採用した定義付けを行っている [1]。

一方、日本語圏の Splog 空間には主に 3 タイプの Splog が存在している [2]。以下にその 3 タイプの名称と各タイプの特徴を挙げる。

(1) アフィリエイト型 Splog

記事内に多数のアンカーテキストを含ませ、訪問ユーザにそのリンクを辿らせることでアフィリエイトとして発生する広告収入を得ることを目的としているブログサイト。

(注1): <http://wikipedia.org/>

(2) コピー＆ペースト(コピペ)型 Splog

話題のホットピックを含んだ記事を他サイトからコピー＆ペーストすることで、サイトアクセスの誘導、アフィリエイトを目的とした順位操作を行う。コピー＆ペーストという単純作業により記事の大量生成が可能といった特徴がある。

(3) ワードサラダ型 Splog

話題のホットピックを含んだキーワードを使用して、文書としては成り立っているのだが文書自体には全く意味がない記事を生成しそれを記事として公開する。記事本文自体は自動で生成されているためこのワードサラダ型もコピー＆ペースト型同様、記事を大量に生成することが可能。

以上2つの要素を踏まえた上で以下に、本論文で用いる Splog 定義を示す。

商品の宣伝や広告、アダルトコンテンツを含んだブログ記事を生成し、本来のブログ目的とは異なるアフィリエイト目的など、ユーザにとって決して有益でないと思われるブログ

以上の定義を本論文で用いる Splog 定義の出発点とする。また、本定義とともに上記で挙げた日本語圏における3タイプの Splog も Splog として扱う。

また、我々は Splog 空間には、(1)万人に共通する Splog 空間と(2)ユーザごとに異なる Splog 空間の2種類が存在すると考えている。このため、Splog フィルタリングには、ほぼ普遍的でありかつ随時更新可能な共通フィルタリング部とユーザごとに異なる Splog フィルタリングが可能な個人適応型フィルタリング部の2つが有効であると考えられる。

そこで、本研究では日本語圏の Splog 空間に関する基礎的な知見の獲得とユーザ適応型のフィルタリングの必要性の検証を目標とし、Splog 空間調査支援システム SplogExplorer を開発した。本ツールは3つのサブシステムから構成されており、それぞれのサブシステムが Splog 空間を調査、分析するための機能を提供する。

本論文の構成は次の通りである。2.では、スパムフィルタリングに関する先行研究をレビューし、本研究における新規性を考察する。3.では、開発したツール SplogExplorer についてその機能と利用方法について述べる。4.では、ユーザ適応型フィルタリングの必要性の検証実験として、SplogExplorer を用いた Splog 空間検証実験を行った。5.では、4.章で行った検証実験についての考察をし報告する。最後に、6.では、本論文のまとめと今後の展開について述べる。

2. スパム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 を取り入れ F 値約 70% の Splog 検知に成功している [3]。使用している SVM の特徴としてはブログサイトの「URL」やそのブログ記事が持つ「アンカーテキスト」を新規特徴として提案しその効果について検証している。また従来のスパムメールフィルタリングなどで使用されている「Words」[4]にも着目しこれら従来の特徴と新規提案特徴とを

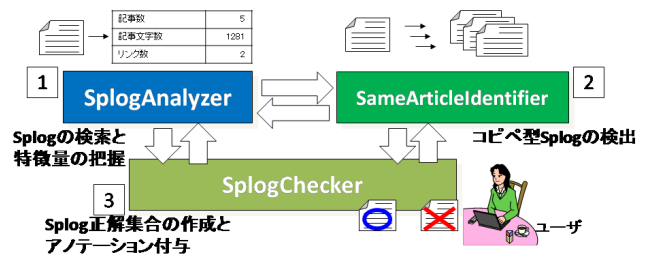


図1 SplogExplorer 全体構成図

組み合わせで使用している。

日本語圏における Splog 研究としては石田の研究 [5] 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っている。ブログサイトにおける外部リンクに着目し高次元のリンク数においては、90%以上もの Splog フィルタリングを可能としている。

以上の先行研究において、それぞれの Splog フィルタリングは高い有効性を示している。また、Kolari の研究 [6] から、日本語圏の Splog 空間と英語圏の Splog 空間とは異なる傾向にある。彼らが提案する 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、共通であるが故にユーザ個人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っている。ユーザ適応型フィルタリングとは、ユーザそれぞれに対応した柔軟なフィルタリングを実装することが可能であるため、ユーザは真に必要とする情報だけを取得することができる。

本論文では、ユーザ適応型フィルタリングに必要な基礎データの収集と分析を行いユーザ適応型フィルタリングの必要性の検証を行う。

3. Splog 空間調査支援ツール:SplogExplorer

本研究では Splog 空間を調査、分析するために支援ツール SplogExplorer を開発した。Splog 空間を調査分析するためには(1)キーワードや数値情報による Splog 空間のサンプリングによる分析と(2)各 Splog3 タイプを検知すること(3)研究に必要な Splog データセットの作成 [7] が必要である。そのため我々が開発した SplogExplorer は以下の3つのサブシステムで構成しており、それぞれのサブシステムが Splog 空間を調査、支援するための種々の機能を提供する。図1に SplogExplorer の全体構成図を示す。

(1) SplogAnalyzer

リンク数や文字数などの特徴量に着目した Splog 空間分析 [8] を様々な視点から行うことが可能なサブシステムである。

(2) SameArticleIdentifier

コピペ型 Splog の検知に特化したサブシステムである。

(3) SplogChecker(SC)

効率的なデータセット作成支援を目的としたサブシステムである。ユーザ固有の Splog 空間を形成、確認することができる。

以下データセットの説明後、評価実験に用いる SC システムに焦点を絞って説明する。

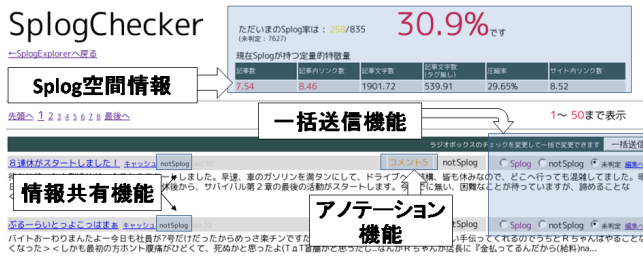


図 2 SplogChecker ユーザインタフェース

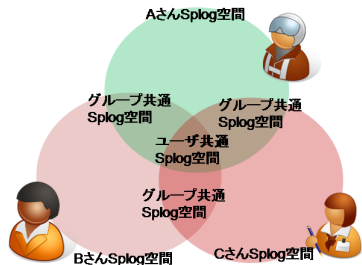


図 3 ユーザ間における Splog 空間イメージ図

3.1 データセット

本論文で使用するデータセットは 2007 年 4 月 30 日、実際に Web 上で公開されたブログ記事 8462 件を使用する。記事 8462 件のサイト数は 5467 サイトである。

3.2 SplogChecker

SplogChecker(SC) システムは Splog データセット作成支援を目的としたシステムであるが、その特徴の 1 つとしてシステム上でユーザ固有の Splog 空間を形成することが可能となっており、ユーザは Splog 空間における様々な情報を得ることができる他、ユーザとの Splog 空間情報共有などが行える。また、研究を進めて行く上で Splog データセット作成という作業は必要不可欠な重要プロセスである [7]。図 2 に SplogChecker システムのユーザインタフェースを示す。SC システムにはデータセット作成支援のために様々な機能が提供されており、その上でユーザ固有の Splog 空間を形成することができる。以下に SC システムが提供する機能を示す。

- (1) Splog フラグ一括送信機能
- (2) アノテーション機能
- (3) 他ユーザとの情報共有機能
- (4) データセット内における Splog 空間情報

4. 評価実験

本論文では以下 2 つの評価実験を行った結果を示す。

- (1) ユーザ固有 Splog 空間検証実験
- (2) Splog 判定基準の検証

我々は Splog 空間というものがある共通した一つの空間ではなく個々のユーザごとに存在していると考えている (図 3)。そのため本論文では、その存在の有無について証明、検証するために SC システムを用いた評価実験として (1) を行った。また、評価実験後に被験者へは個別にインタビューを行い、更なる Splog 空間分析として (2) を行う。

以下、実験環境および実験方法について説明した後、結果と考

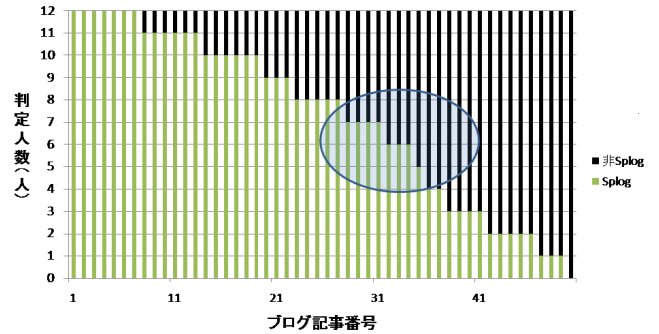


図 4 共通記事 50 件の Splog:非 Splog 判定割合

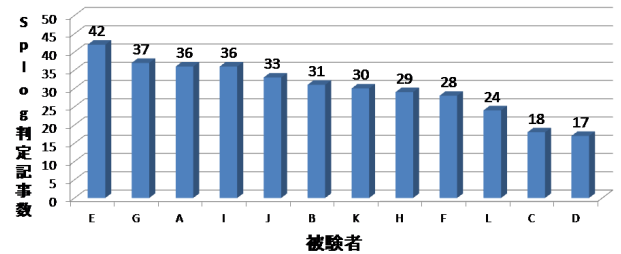


図 5 被験者ごとによる Splog 判定記事数の分布

察を述べる。

4.1 評価実験環境および実験方法

- 被験者

実験を行う被験者は日常的に Web を利用している工学系大学生 12 名で、男女比は 11:1、年齢は全員 20 代である。また、被験者へは事前に本論文の Splog 定義と実際の Splog 例についての説明を 10 分程度行った。

- 評価実験に用いるブログ記事

被験者に判定してもらったブログ記事は 50 件で共通の範囲を設ける。また、共通記事 50 件の内容は我々が事前に選定し、その内訳を Splog:非 Splog=35:15 とした。この 50 件の共通記事内には本論文で示した Splog 例 (アフィリエイト型、コピペ型、ワードサラダ型) も含まれている。

- 実験方法

50 件の共通ブログ記事に対して被験者 12 名が SC システムを実際に使用して Splog か非 Splog かのフラグ判定を記事に対して付与する。

4.2 ユーザ固有 Splog 空間の検証

評価実験の結果としてまず、共通記事 50 件がそれぞれのどのような割合で被験者 12 人に判定されたのかを図 4 に示す。縦軸は被験者の判定人数をとり、横軸には 50 件のブログ記事をとっている。被験者は 12 名で行っていることから縦軸の最大は「12」となっている。

第 2 の実験結果として各被験者ごとによる Splog 判定の記事数の分布を図 5 に示す。縦軸にはユーザが Splog と判定した数をとっているため最大は 50 となり、横軸には被験者 12 人をそれぞれとっている。またこの図において、左から Splog を多く判定した被験者でソートしているため右に行くほど Splog 判定数が少ない被験者となっている。

第 3 の実験結果としては、評価実験で用いた 50 件のテスト

表 1 特徴的な結果を示した記事 12 件とその判定マトリクス

記事整理 No.	BL	A	B	C	D	E	F	G	H	I	J	K	L
1	1	0	1	0	0	1	1	1	1	0	1	0	0
2	1	1	1	0	0	0	0	1	0	1	1	0	1
3	1	1	0	1	0	1	0	1	0	0	1	0	1
4	1	1	1	0	0	1	0	0	0	1	0	1	0
5	1	0	0	0	0	1	0	1	0	1	0	1	0
6	1	1	0	0	0	1	0	0	0	0	0	1	0
7	1	1	0	0	0	0	0	1	0	1	0	0	0
8	0	0	1	1	1	1	1	1	1	1	1	1	1
9	0	1	1	0	0	1	1	1	1	1	1	1	1
10	0	1	1	0	0	1	0	1	1	1	1	1	1
11	0	1	1	0	0	1	1	1	0	1	0	1	0
12	0	1	0	0	1	1	1	0	0	1	1	1	0
Splog 判定数	7	9	7	2	2	10	5	9	4	9	7	8	5

データにはあらかじめ我々がベースラインとなる判定を行っている。その中で

- (1) 判定が大きく割れた記事
 - (2) ベースラインは Splog であるにも関わらず、被験者の判定の多数決が非 Splog となった記事
 - (3) ベースラインが非 Splog であるにも関わらず、被験者の判定の多数決が Splog となった記事
- が存在した。表 1 にそれら記事における被験者らの判定マトリクスを示す。

この表で「BL」がベースラインで我々が判定した基準を示しており、ベースラインを基準に上に Splog, 下に非 Splog が来るようにソートされている。

4.3 インタビューによる Splog 判定基準の検証

さらに我々は表 1 で示したテスト記事に対して 12 人の被験者のうち 8 人の被験者に個別にインタビューを行った。これは被験者が一体どのような Splog 判定基準を持って判定したのかを分析するためであり、被験者の判定基準から Splog 空間の存在を検証するために行った。行ったインタビューは 2 項目であり、それぞれ (1) 被験者とブログに関する一般的な質問を行った後 (2) 表 1 のテスト記事 12 件に対してなぜその判定をしたかを聞いた。

まず、(1) の質問内容の一部を以下に示す。

- ブログについてどう思うか?
- Splog についてどう思うか?
- ブログサイトを自身で所持しているか?
- ブログをどういった時に活用しているか?

以上の質問を行った結果、被験者のブログに対する傾向をまとめたものを下記に示す。

- ブログ記事の閲覧は検索エンジンによる検索結果からの訪問が大半である
 - 実際 Splog に出くわすことはよくあり、それが不快に感じることはある

ということが傾向としてよく見られた。また、自身でブログサイトを所持しているのは 8 人中 2 人であった。

このインタビューからわかったことは、今回の被験者群は定期的に購読するようなブログサイトは少ないものの情報を得る手

表 2 テスト記事 12 件の特徴と概要

記事整理 No.	リンク	画像	文字	バナー	キーワード
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					

表 4 各被験者における Splog 判定タイプ分類表

被験者	内容	リンク	キーワード	見た目
A				
D				
E				
F				
H				
I				
K				
L				

段としてはブログを活用していることがわかった。また、Splog に対しても知っている被験者は大半でかつそれを不快に感じる人がほとんどであることがわかった。

次に、(2) のインタビュー結果としてまず、テスト記事で使用した 12 件のテスト記事の特徴をまとめた結果を表 2 に示す。それぞれの項目は判定者が主観で判定した特徴を表している。例えば「リンク」ならば記事中に多数のリンクを含んだ記事で構成されているならばリンクに「 」が付けられる。同様に「文字」「画像」「バナー」が多いものには「 」が付けられる。最後に「キーワード」だがこれは Splog として印象を受けるキーワード (例えば「儲かる」、「アダルト」など [7]) が多く含まれている場合に「 」が付けられる。

インタビューはこれら 12 件の記事に対してそれぞれ 8 人の被験者に対して個別に実施した。インタビュー内容は「被験者がなぜそのような判定をしたか」である。各記事に対する被験者のインタビュー結果を記事の簡単な説明と共にまとめたものを表 3 に示す。最後に、このインタビューデータ全てを通して被験者の分析を行い、被験者の Splog 判定タイプを割り出した。表 4 にその結果を示す。インタビューの分析結果から項目にはそれぞれ「内容」「リンク」「キーワード」「見た目」を設けた。それぞれの項目に関して判定の際重視する項目には「 」を付与する (特に重視する項目がある場合は「 」とする)。

次の章で、以上 2 つの評価実験

- (1) ユーザ固有 Splog 空間検証実験
- (2) Splog 判定基準の検証

表 3 テスト記事 12 件に対するインタビュー結果のまとめとその概要

記事整理 No	概要	Splog と判定した被験者	非 Splog と判定した被験者
1	本の紹介サイト	本の写真が多くあり、リンクもしてあることからアフィリエイトサイトだと思った。(E, H, I)	本の紹介情報としてはしっかりしている。アフィリエイトサイトには見えなかった。(A, D, F, K, L)
2	マグカップの紹介サイト	「ロコミ」というキーワードが怪しい。商品の羅列のためアフィリエイトだと感じた。(A, K, L)	商品に対するレビューがしっかりしている。リンクは多かったが怪しいリンク先でなさそうなのでアフィリエイトだとは思わなかった。(D, E, F, H, I)
3	求人情報サイト	ブログで求人情報を紹介する事自体怪しい。このサイトのブロガーの意図がよくわからない。(E, K, L)	求人情報としてはしっかりしている。ぱっと見は普通のブログに見えた。(D, F, H)
4	カラオケ情報 (メルマガに関するサイト)	メルマガという事自体怪しい。また、メルマガを載せているだけでそもそもこれをした理由がわからない。(D, E, L)	ぱっと見情報もあることから普通のブログであると判断した。嬉しい情報であったから非 Splog にした。(F, H, I, K)
5	様々な病院を紹介しているサイト	全体的に文字ばかりでごちゃごちゃしていてわかりづらい。タイトルが怪しく感じた。(A, D, E)	「病院」というフレーズから信頼できる情報だと信じた。情報としてはしっかりしている。(H, I, L)
6	某アイドル紹介サイト	とりあえず見づらい。意味がわからない (と感じた人が大半)。女性の写真が多いのは怪しいサイトに見えてしまう。(D, E, L)	内容自体は非 Splog ではあるが、見た目はアイドルの写真ばかりで少し見づらい感じにも思えた。画像が多くてぱっと見はよくわからなかったの、内容を見て判断した。(A, F, H, I, K)
7	エステ情報紹介サイト	キーワードに「激安」「無料体験」などが見えたので Splog と判断。(A, L)	情報としてはしっかりしている。リンクが少ないためアフィリエイトではないと判断した。(D, E, H, K)
8	ネットビジネス紹介サイト	すぐに Splog と判断できた。アフィリエイト情報を紹介しているのは明らかだったので即 Splog 判断。(D, I, K) キーワードが怪しい。(A, E, F, H)	アフィリエイトに興味があるので非 Splog にした。必要な状況下にあるときに閲覧できれば情報としては嬉しい。(L)
9	他のブログ記事へのリンク集	見た目のリンクの多さから Splog と判定した。(A, H) タイトルと記事内容に統一性がないため怪しいと感じた。リンク集ではあるが単なる他のサイトからのコピペ記事だと思った。(D, E, K, L)	内容に害は感じられなかったし、アフィリエイトとしてのサイトではないと判断した。(F)
10	お金の稼ぎ方を紹介しているサイト	あきらかにアフィリエイトである。リンクが怪しすぎる。キーワードも「モテル」「稼ぐ」など多数見受けられる。(A, E, I, L)	紹介している情報としてはしっかり紹介している。情報としては嬉しい情報が多い。(F, H)
11	記事整理 No9 と同様のサイト	リンクが多い。記事のリンク集に統一性が見られない。(E) 情報としてもそれほど重要な情報ではなさそう。(A, D, H, L)	ぱっと見、日記を書いている普通のブログサイトだと思った。リンクは多いが怪しいリンクではなさそう。(F, K)
12	お気に入りサイト一覧集	リンクのみで構成させているのはいかにも怪しい。(A, D, E, L) タイトルにアンカーを貼っているだけで内容としてもよくわからない。(F, H)	自分のお気に入りを集めているだけであり、アフィリエイトリンクではないと判断した。情報として害のない情報であると判断した。(I, K)

についての考察を行う。

5. 評価実験考察

この章では 4. 章で行った評価実験に関する考察を各実験ごとに行い、ユーザ適応型フィルタリングの必要性についてまとめる。

5.1 ユーザ固有 Splog 空間の検証に関する考察

まず、図 4 から、ユーザ間において明確に判定が一致したブログ記事が存在していることがわかる。つまり被験者 12 人全員が「Splog」または「非 Splog」と判定したブログ記事が図の両端に表れている。このような判定をされたブログ記事はユーザ間で定義に差異がないものと考えられる。つまり、ユーザ共通の定義と考えてよいブログ記事である。被験者 12 人全員が「Splog」と判定したブログ記事は全部で 7 件あり、アフィリエイト、アダルト系、出会い系のブログサイトであった。また、逆

に 12 人全員が「非 Splog」と判定したブログ記事は「日記」であり、全部で 1 件であった。

次に、判定が割れているブログ記事も存在しているということがわかる。図 4 の中央付近の円で囲んだ部分がそれに該当し、Splog:非 Splog=6:6 となっている近傍のブログ記事群のことである。評価実験では事前に Splog 定義と Splog 例については被験者に説明してある。そのため、判定が割れたブログ記事はユーザ間において Splog に対する価値観や認識に差異があり、ユーザ固有に Splog 定義が存在しているという可能性を示している。判定が 50%に割れたブログ記事は全部で 3 件あり、ブログ記事はアフィリエイト型であった。アフィリエイト型にも関わらずこのような結果となった原因を以下に示す。

- (1) アフィリエイトであったにも関わらず、そのユーザにとっては真に必要な情報であったから
- (2) 被験者は判定したブログ記事をアフィリエイト型だと

見抜けなかった(ブログサイトに騙された)

(1) はたとえアフィリエイトブログ記事であっても、全てのユーザにとって、その記事が不要であるとは限らないことを示唆している。自分が得たい情報が得られるのであれば、そのユーザにとっては有用なブログ記事となる。このため、一律にアフィリエイト記事をフィルタリングして排除してしまうことには問題がある。逆に、ほとんどのユーザが通常のブログ記事であると判断したにも関わらず、あるユーザにとってはそれが Splog になるという事例もある。その事例として表 3 の記事整理 No.8 が該当する。判定割合は Splog:非 Splog=11:1 である。

(2) は、ユーザがブログ記事を読んだだけではアフィリエイト目的であることが分からない場合があることを示している。このような場合には、Splog の特徴量 [2] を用いることによってユーザにアフィリエイトの危険があることを通知する機能が有効になる。

また、(2) 記事の例としては、表 3 の記事整理 No.5 が該当する。判定割合は Splog:非 Splog=4:8 である。

この記事では 2 名の被験者 (F, K) が騙されていたという意見を述べていた。他の非 Splog 判定者に騙されたという意見はなかったが、このように被験者が多数決で非 Splog である記事を Splog だと間違えて判定するというケースも存在した。

また、図 5 からはユーザ間での Splog に対する評価が異なっているということがわかる。共通記事 50 件のうちもっとも多く Splog と判定した被験者は 42 件もの Splog があると判定しているのに対して、最小では 17 件しか Splog としてのブログ記事はないと判定している被験者もいる。

最後に表 1 についての考察を行う。表 1 で Splog 空間が利用者すべてに共通であるならば、ベースラインを境にソーティングした上部が「1」となり、下部が「0」となるはずであるが、この表からもわかるように判定には被験者ごとによりかなりの差が見られており、更にベースラインと反した判定の記事も存在することから、それぞれに固有の Splog 空間が存在していることがわかる。

5.2 Splog 判定基準の検証に関する考察

まず表 3 からわかる考察として判定が分かれた被験者それぞれによって意見は異なり、それぞれが自分の主観的意見に沿った判定を行っているということがわかる。非 Splog と判定している意見の傾向としては「情報として必要である」という意見がやはり強く、逆に Splog と判定している意見の傾向としては「アフィリエイトだと思った」「怪しいサイトに見える」など、自分にとってのマイナス要素がある場合には Splog に判定するという傾向が見られた。

また表 3 と表 4 の Splog 判定タイプをまとめた結果から、被験者が Splog 判定を行う際の基準となる傾向として、ほとんどの被験者はまずブログサイトを訪問した際、そのブログサイトの見た目で判断を行うという傾向が見られる。そして、見た目で判断ができない場合に初めて内容をじっくり読みはじめるといった傾向があることがわかった。また、被験者は判定を行う際に自分の主観的意見を強く尊重するため、自分の Splog 基準を含んでいないブログサイトが実際には Splog であった場合でも

非 Splog の判定を行う傾向がある。

このように Splog を判定する際にも、ユーザそれぞれにおいて判定に差異が存在し、それぞれが主観的な基準を持ち合わせて判定を行っている。そのため Splog 空間にもユーザごとで差異が発生している。

以上これら 2 つの評価 (1) ユーザ固有の Splog 空間検証と (2) Splog 判定基準の検証結果からユーザ適応型フィルタリングの必要性について考察することができた。そしてこれらの実験結果から、Splog の有効なフィルタリングには

- (1) ユーザに依存しない共通的なフィルタリング
- (2) ユーザに固有な個人適合可能なフィルタリング

の 2 種類が必要であることがわかる。

6. おわりに

本研究では、Splog 空間の性質、価値を明らかにするために、複数ユーザが同一ブログ空間内にアクセスし、個々のブログ記事に対して各ユーザの Splog 判定結果を付与することができ SplogChecker システムを開発した。12 名の被験者を用いた評価実験より、ユーザに依存しない Splog 空間が存在すると同時に個々のユーザの特性に依存する Splog 空間も存在することが確認できた。また、ユーザの Splog 判定基準もそれぞれが異なった主観的意見を持ち合わせていることからユーザ固有の Splog 空間の存在を証明することができた。今後は、本実験で得られた知見を元に Splog の効率的なフィルタリング機能および、個人個人に適合可能なフィルタリング、そして、ブログ記事が Splog である可能性が高い場合にはユーザに通知するアラート機能を実現していく予定である。

文 献

- [1] Pranam Kolari, Akshay Java, and Tim Finin. Characterizing the splogosphere. *Proceedings of the 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 15th, World Wide Web Conference*, May 2006.
- [2] 芳中 隆幸, 福原 知宏, 増田 英孝, and 中川裕志. Splog 空間における定量的調査支援システムの開発とその評価. 第 22 回人工知能大会全国大会, June 2008. 1E1-01.
- [3] 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.
- [4] Drucker H., Wu D, and Vapnik V. Support vector machines for spam categorization. *IEEE-NN*, pages 1048–1054, 1999.
- [5] 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*, Nov 2007. 5B.
- [6] Pranam Kolari, Akshay Java, Tim Finin, Tim Oates, and Anupam Joshi. Detecting spam blogs: A machine learning approach. *Ph.D. Dissertation*, Dec 2007.
- [7] 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, and 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・類似化・データセット作成. *DEWS2008-第 19 回データ工学ワークショップ*, March 2008.
- [8] 芳中 隆幸, 福原 知宏, 増田 英孝, and 中川裕志. スパムブログに関する定量的調査支援ツールの開発. 情報処理学会第 70 回全国大会, March 2008. 5J-7.