1E1-01

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

Development of An Analysis Support System of Splogosphere and its Evaluation

芳中 隆幸[*1] 福原 知宏[*2] 増田 英孝[*1] 中川 裕志[*3]
Takayuki Yoshinaka Tomohiro Fukuhara Hidetaka Masuda Hiroshi Nakagawa

[*1]東京電機大学未来科学部
School of Science and Techonology for Future Life, Tokyo Denki University

[*2]東京大学人工物工学研究センター
Research into Artifacts, Center for Engineering, The University of Tokyo

[*3]東京大学情報基盤センター
Infomation Technology Center, The University of Tokyo

Spam blogs (Splogs) become a one of big issues on the Web because splogs confuse search engines so that spam pages are ranked highly. Investigating splogsphere is necessary for filtering out splogs efficiently. In this paper, we propose a system called SplogExplorer for investigating the splogosphere. By using this system, researchers can find various types of splogs easily. Evaluation results of the system are also described.

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, 多くの人々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 (スパムブログ,Splog(スプログ)) 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.

Kolari らは英語圏の Splog 空間について調査を行っている [1] が, 日本語圏の Splog 空間は英語圏とは異なる傾向にある. また, 日々新たな種類の Splog が出現し, いたちごっこの状態が続いており, 効果的な Splog フィルタリングの実現には,Splog 空間についての十分な知見が必要である [2]. また, 我々は万人に共通する Splog 空間と, 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, ほぼ普遍的であるが随時更新可能なフィルタリング部と個人適応型フィルタリング部が必要となる.

そこで, 本稿では日本語圏の Splog 空間に関する基礎的な知見の獲得と, ユーザ固有のフィルタリングの必要性の検証を目標とし, 定量的 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 空間を分析するために種々の機能を提供する.

本論文の構成は次の通りである. 2. では, 本稿で用いる Splog 定義と特徴量について考察し, それら 2 つの関係性について述べる. 3. では, 開発した SplogExplorer の機能とその利用性について述べる. 4. では,SplogExplorer を用いた評価実験を行い,Splog 定義に関する考察を行う. 最後に,5. では, 本稿のまとめと今後の展開について述べる.

2. Splog と特徴量

この章では日本語圏における Splog 定義がどういったものかについて考察し, 定義付けを行うことで,Splog の実態を明確にする. また本稿で着目する特徴量についても述べる.

連絡先: 芳中 隆幸, 東京電機大学未来科学部情報メディア学科, 東京都千代田区神田錦町 2 丁目 2 番地, 03-5280-3281 ext 2843, 03-5280-3592, yoshinaka@cdl.im.dendai.ac.jp

2.1 日本語圏における Splog とその定義

Kolari らは研究における Splog 定義を Wikipedia[*1]から採用している [3] が, 本稿では, 独自の Splog 定義を提案することにする. これは研究を進めるにおいて,Splog という対象を明確化するという狙いがある. 以下, 本稿で用いる Splog 定義である.

> 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

また, この本 Splog 定義から考えられる, 実際の日本語圏における Splog 空間に存在する Splog 例を以下に示す.

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
2. コピー&ペースト (コピペ) 型 Splog
   話題のホット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, サイトアクセスの効率化, 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.

この Splog 定義と Splog 例を本稿で用いる共通の Splog 定義とし研究を進めていくこととする.

[*1] http://wikipedia.org

表 1: Splog における特徴量

特徴量	説明
記事数	ブログサイトが一日に書く記事数
記事内リンク数	ブログ記事内のみでの外部リンク数
サイト内リンク数	ブログサイト全体での外部リンク数
全文字数	ブログ記事内の文字数
タグ無し文字数	ブログ記事内のタグを除いた文字数
圧縮率	タグ無し文字数/全文字数

2.2 特徴量

本稿で着目する特徴量を表 1 に示す. これら特徴量はブログサイトが必ず所持している情報であり, これら特徴量を Splog 空間を分析するためのパラメータとして用いる. ここでは, 特徴量と Splog との関係性について考える.

「記事数」とはブログサイトが一日に書く記事数のことであり, 記事を大量生成するコピペ型 Splog やワードサラダ型 Splog の分析に基づく特徴量である. これは Splog と notSplog の記事数に着目した場合, 記事数を多く書いているブログサイトは傾向として Splog に近いというが分析でき, 特徴量の差異を利用することで Splog 検知の実現を可能とする.

「記事内リンク数」と「サイト内リンク数」はブログサイトが持つリンクに着目した特徴量であり, 多数のアンカーテキストを含んでいるアフィリエイト型 Splog に基づく特徴量である. アフィリエイト型 Splog には当然これらの特徴量が多く見られる [4] ことから, 「記事数」同様に Splog と notSplog との差異に着目することで, 傾向として Splog を検知しやすくなる. 「全文字数」,「タグ無し文字数」,「圧縮率」はブログサイトの文字数の関係に着目した特徴量で, タグとの比率を分析することにより, アフィリエイト型 Splog の傾向が把握しやすくなる.

このように各 Splog 例との関係性を用いることにより, 本稿ではこれら特徴量に着目した研究を進めて行くことにする.

3. 調査支援システム:SplogExplorer

本稿では Splog 空間を調査, 分析するために支援システム SplogExplorer を開発した. この SplogExplorer は以下 3 つのサブシステムで構成されており, それぞれのサブシステムが Splog に対し種々必要な機能を提供する.

1. SplogAnalyzer
2. SameArticleIdentifier
3. SplogChecker

図 1 に SplogExplorer の全体図, 図 2 にユーザインタフェースを示す.

以下本稿で用いているデータセットの説明をした後, それぞれのサブシステムについて詳しく説明する.

3.1 データセット

本稿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 件に相当するサイト数は 5467 サイトである.

3.2 SplogAnalyzer

SplogAnalyzer(SA) システムはデータセット内における分析ツールとしてブログ検索を提供しているサブシステムである. この SA システムの特徴の一つとして, 検索時 (図 3) に特徴量をユーザに対し提示するという機能が備わっている. そのためユーザは Splog 空間における特徴量を容易に分析, 把握できる他, Splog と notSplog における特徴量の差異も知見として得ることができるため, 特徴量に着目した Splog 空間分析が可能となっている. また,SA システムには様々な検索機能が備えられおり, キーワードによる検索の他, 特徴量による検索や, ブログサイトのドメインごとにおける検索機能などが備わっているため, 多種多様な検索から Splog 空間を分析することが可能である.

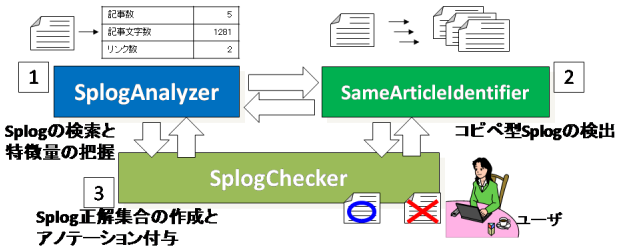

図 1: SplogExplorer 全体構成図

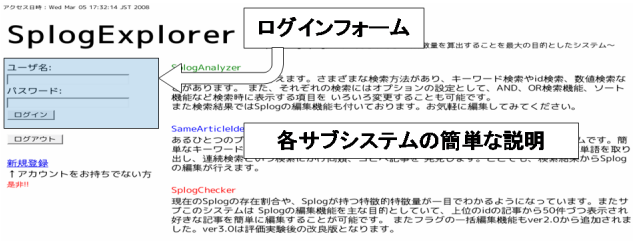

図 2: SplogExplorer ユーザインタフェース

3.3 SameArticleIdentifer

SameArticleIdentifer(SAI) システムは, 他のブログサイトの記事を無断引用するコピー&ペースト型 (コピペ型)Splog を検出するために特化したシステムである. ここで SAI システムにおいてコピペ型 Splog を検出するためのアルゴリズムを図 4 に示す.

まず SAI システムでは, コピペの元となる記事を 1 記事選定する. 記事を選定したらその記事に対して解析を行う. 本稿では解析器として Sen[*2], また専門用語抽出として, TermExtract[*3] を使用している. 実際にシステムで記事を解析した結果を図 5 に示す.

解析結果では, 元となっている記事の情報としてタイトル, 本文情報が提示される他, 記事が持つ形態素の情報も提示される. 抽出する形態素は主に名詞（複合語）, 形容詞, 未知語である. また,TF-IDF 法による単語への重み付けによる抽出も行っている. 重み付けを行うことにより, 記事内での特徴的な語を抽出することができ, その語の話題性に着目することで, 記事内容として同様の記事を抽出することも可能としている.

解析を行った後, これらの記事解析データを元にしてコピペ型 Splog の検知を行う. 記事解析データから複数の単語を選定し, その単語群による連続的な AND 検索をかけることにより, 全体の記事集合を縮小させていく. 記事集合を縮小させる

*2 https://sen.dev.java.net/
*3 http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html

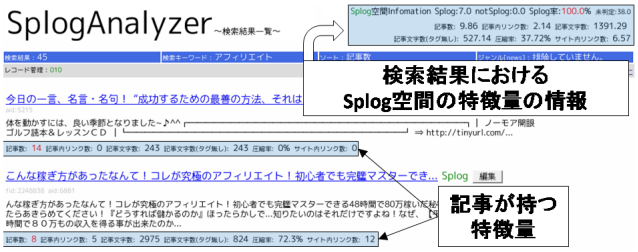

図 3: SplogAnalyzer ユーザインタフェース

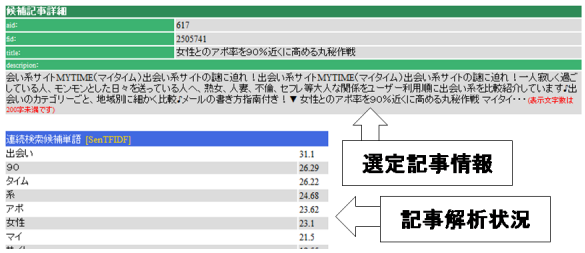

図 5: SameArticleIdentifier による記事解析結果

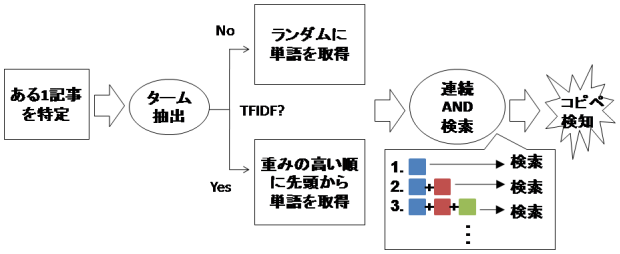

図 4: コピペ型システム検知アルゴリズム

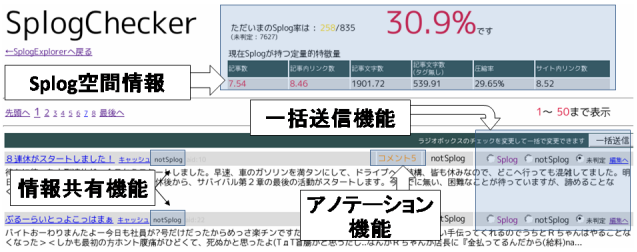

図 6: SplogChecker ユーザインタフェース

ことにより, 最終的には元になっている記事とそれと同様の記事が残るため, コピペ型 Splog の検知を可能としているのがこの SAI システムの検知プロセスである.

SAI システムの特徴として部分的なコピペに対応可能という点が挙げられる. SAI システムでは単語を選定する際, 隣接する単語を取り出しているため単純な文字列マッチングでは不可能な, この点に対応することが可能となっている. また, 現状の SAI システムによる検知プロセスではコピペ型 Splog として可能性のある記事を検知するだけであるため. 今後より精度の高い検知プロセスを実現することも必要であると考える.

3.4 SplogChecker

SplogChecker(SC) システムは効率的な Splog のデータセット作成支援を目的としたシステムである. 研究を行う上で,Splog データセット作成という作業は必要不可欠かつ重要なプロセスである [5]. 図 6 に SplogChecker システムのユーザインタフェースを示す.

SA システムや SAI システムでは検索結果から Splog フラグを付加できるのが, 本 SC システムでは Splog フラグの付加を一括で行うことができ, 効率の良い Splog フラグ編集が可能となっている. その他にも 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いる. 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

3.4.1 Splog フラグ一括送信機能

複数記事における判定フラグ情報を一括で送信できる機能である. 1 記事ごとにフラグを情報を送らなくて良いため効率の良いフラグ判定を行うことが可能となっている.

3.4.2 アノテーション機能

記事に対しコメントを付加できる機能である. 気になる記事, 判定が困難な記事に対しその旨を書き留めておくことが可能で, ユーザは SplogExplorer トップ画面 (図 2) で実際にコメントした記事を確認することができるためどの記事に対して, どのようなコメントをしたかが一目でわかるようになっている.

3.4.3 他ユーザとの情報共有機能

システム利用者が他ユーザと Splog 判定情報を共有できる機能である. この機能により, ユーザが判定に手間取っている, 困惑している時に他のユーザの情報を共有することで, 判定を効率的かつ正確に行うことを可能としている.

3.4.4 データセット内における Splog 空間情報

SA システムでは検索結果ににおいて Splog 空間が持つ特徴量の統計情報を提示しているが, SC システムにおいては, データセット内全てにおける Splog 空間情報を一目で把握することができる. またデータセット内に現在どれくらいの Splog が存在しているかの割合情報も提示するためリアルタイムに Splog 空間情報を実感することが可能となっている.

4. SplogChecker システムを用いた評価実験

SplogChecker システムを用いた評価実験を行った. 我々は Splog 定義というものが共通した一つの定義ではなくユーザ固有で存在していると考えている. そのため本稿ではその存在の有無について評価を行う.

以下, 実験環境について説明した後, 結果と考察を述べる.

4.1 評価実験環境

- 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学生 12 名. 男女比は 11:1 で, 年齢は全員 20 代である. また, 被験者へは事前に本稿 Splog 定義と Splog 例についての説明をしておく.
- 評価実験に用いるブログ記事
  被験者に判定してもらうブログ記事 50 件は全部で 100 件, その内 50 件は被験者ごとによる範囲で残り 50 件は被験者

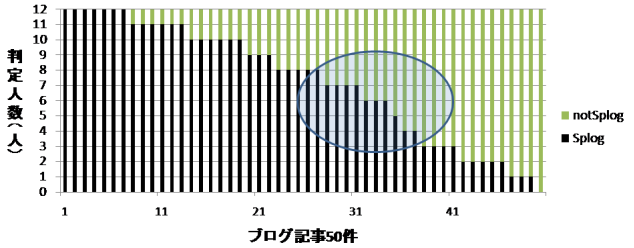

図 7: 共通記事 50 件の Splog:notSplog 判定割合

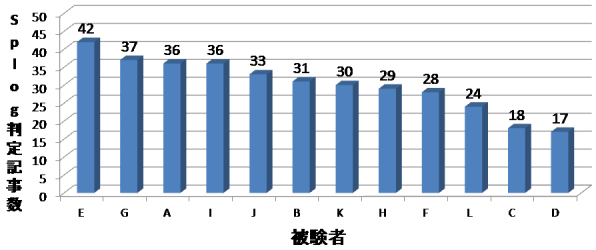

図 8: 被験者ごとによる Splog 判定記事数の分布

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:notSplog=35:15 とした．この 50 件の共通記事内には本稿で使用する 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．

4.2 評価結果

評価の結果としてまず, 共通記事 50 件がそれぞれどのような割合で被験者 12 人に判定されたのかを図 7 に示す．この図 7 からわかることとして, まず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「notSplog」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というの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と考えられる．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「notSplog」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, 判定が完璧に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 7 の中央付近の部分がそれに該当し,Splog:notSplog=6:6 となっているブログ記事のことである．これらのブログ記事はユーザ間において Splog に対する価値観に差異があると考えられ, 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．判定が割れたブログ記事は全部で 3 件あり, それらの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被験者ごとで判定が割れた原因であると思われる項目を以下に示す．

1. アフィリエイトであったにも関わらず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）
3. Splog に関する知識不足

これらを解消する手段として, やはり Splog に関する認知, 知見の獲得が必要であると考えられる．また, 判定が割れたブログ記事において被験者で議論をするということも挙げられる．

2 つ目の評価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 8 に示す．図 8 からもわかるように, やはりユーザ間での Splog に対する価値観は異なっている．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．

以上 2 つの評価結果からユーザ固有での Splog 定義実現は効果的かつ有効な手段であると考えられる．また, 被験者間で判定に差異のでない記事も存在することから, ユーザ共通の Splog 定義を実現し, その付属定義としてユーザ固有の定義を作成することが実現に向けた最善手段であると考える．

5. おわりに

本稿では Splog の明確化を行った上で, Splog 空間の基礎的知見を得るための調査支援システム SplogExplorer を開発した．提案システムは 3 つのサブシステムで構成されており, それぞれの主な支援機能として

1. 検索による Splog 空間の特徴量の提示
2. コピー&ペースト型 Splog の検知
3. データセット作成の効率化支援

があり, これらのサブシステムが Splog 空間を定量的に調査支援する．評価実験では,Splog 定義がユーザ固有で存在することを証明することで, ユーザ適応型のフィルタリングを実現することが有効な手段であるということを証明した．

今後の課題としては, まずデータセットの拡張を行うことであり, 現在 1 日分のブログ記事を使用しているが, 1 週間,1ヶ月と期間を拡大することにより, 新たな Splog 空間の在り方を発見できると考えられる．展開としては, 特徴量の有効性の証明として機械学習への適応による精度調査等を検討している．

参考文献

[1] Pranam Kolari, Akshay Java, Tim Finin, Tim Oates, and Anupam Joshi. Detecting spam blogs: A machine learning approach. *Ph.D. Dissertation*, Dec 2007.

[2] 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.

[3] 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.

[4] 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.

[5] 佐藤 有記, 宇津呂 武仁, 福原 知宏, 河田 容英, and 神門典子. 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. **情報処理学会第 *70* 回全国大会**, March 2008. 5J-6.