ブログタイトルに注目したsplogger の判別手法

竹田 隆治 †

† 総合研究大学院大学

著者はこれまで複製文字列検知によるブログエントリ単位での splog フィルタリング手法の研究を行ってきたが，splog を特定するためには，よりシンプルで適切な特徴量があることということを示す。実質的に splogger と呼ばれるユーザがほとんどの splog を生成しており，ブログエントリ単位ではなくブログ単位でこの splogger 判別を行うことについて述べる．

Splogger detection method focusing on the blog title

Takeda Takaharu †

† The Graduate University for Advanced Studies

We have studied the splog filtering technique by the blog entry based on copy string detection. We would like to show simple and appropriate characteristic to detection the splog in this paper. And also, we would discuss to detect a splogger who generates most splog.

1 はじめに

ブログは個人の日記としての情報が記されており，商品やサービスに関してマーケティングを行う上で有益なメディアとして [1] 注目を集めている．一方でブログコンテンツの作成は容易で，ブログの本来の目的に反した splog と呼ばれるスパムコンテンツが存在している．splog は情報検索品質に著しい悪影響を与え，web アーカイブ資源を浪費させるという問題を引き起こしている．

この splog の原因として，splogger と呼ばれる，splog のみを生成し続けるユーザが存在する．本稿では splogger が指す範囲として，splog しか生成しないユーザ（自動生成ツール）と，それに加えて，それを仕込んだ悪意ある人間のことも含めて splogger と呼ぶ．通常の人間の blogger が splog を生成することはほとんどない．実質的に，splog はほぼそのすべてがこうした splogger によって生成されているため，こうした splogger を特定することで容易に splog フィルタリングが可能になる．ブログユーザの区別は URL から簡単にできるため，splogger が特定できれば，その splogger が生成したコンテンツは，内容を確認することなくすべて splog とすることができる．

著者は，複製文字列検知に基づく [6] ブログエントリ単位でのフィルタリング手法の研究を行ってきたが本稿では，このような splogger の特定に関する 1 手法の議論を行う．今回は単独のブログコンテンツとしての splog を検出するのではなく，それを大量に作り続けている splogger の検出を行うために有効な特徴量の議論であるということに注意していただきたい．

splog の本文はその時々によって内容が変わる場合が多い．著者の分類 [6] の通り，(その時度によって変動する）外部コンテンツをコピーしてくるため，内容の傾向は一定ではない．しかしながら，ブログのタイトル及び説明文は基本的に固定であり変わることはない．加えて，splog のタイトルには非常に特徴的な傾向がある．このような，メタデータというべきタイトル，及び説明文を特徴量としたフィルタリングの可能性を示す．

2 関連研究

2.1 フィルタリング手法

splog フィルタリングは，基本的には，検索エンジンの検索上位に位置することを狙った不自然なコンテンツに顕著な傾向に着目することでフィル

タリングを行おうとする．

splog フィルタリングは，リンク解析による手法とコンテンツ解析による手法に大別できる．Kolari らは，英語のブログを対象に splog フィルタリングの先駆的研究を行った [3]．特定種類の単語の割合を特徴ベクトルとして表現し，SVM(Support Vector Machines) などの機械学習の手法を用いて判別を行うコンテンツ解析手法を提案している．

一方，リンク解析を用いた手法として，石田 [2]，Lin ら [7] は，いわゆるリンクファームと呼ばれる大量の不自然な被リンク群に注目した splog（群）の特定手法を提案した．どちらも基本的に同じアイデアでリンク構造からの splog 検出を試みている．

日本語のブログを対象とした研究も始められているが，英語の splog フィルタリングのように文書を特徴ベクトルとして表現した分類問題とは異なるアプローチがとられている．海外の研究では，正解データを用いた機械学習が用いられることが多いのに対し，日本語ブログの研究では，unsupervised な手法を用いることが多い．

著者ら [4, 6] は複数のブログ間に頻出する文字列に着目した．splog は他のコンテンツを部分的または全体のコピーを繰り返して生成されるため，このコピーを検知することで splog の特定が可能になるという考えである．

本稿で提案するフィルタリング手法は，splog の正解データに基づく機械学習による手法である．splog に特有の傾向として，タイトル，説明文という特徴量に注目する．

2.2 評価用データ

筆者らが知る限り，splog フィルタリングの評価用データには，まだ，標準的なものが整備されておらず，splog フィルタリング研究には様々なデータが用いられている．

Lin は [7] TREC Blog Track 2006 を対象とし，ラベル付けした 9200 件のブログエントリ中から，さらに 800 件ずつ blog と splog をサンプリングしている．

Kolari ら [3] はテクノラティ[1]の検索結果からデータセットを構築した．このデータセットも，blog と splog を同数ずつ作製している．blog と splog の比率は，必ずしも実データの比率とは一致していない．

[1] http://technorati.com/

今回我々も独自にコーパスの構築を行ったが，他の研究と違う点としては，ランダムサンプリングによるコーパスであり，実際の BlogSphere の縮図に近づけることを意識したデータであるという点である．

3 評価用データ

3.1 splog の定義

本稿での splog の定義は文献 [6] と同じである．

> 一つ以上の何らかの他のコンテンツを部分的あるいは全てコピーし，それらを連結して生成するブログエントリ

ただし，テンプレート，定型文，単語辞書など，明示的にその存在が明らかではない非公開のコンテンツを用いて作られた語やフレーズもコピーと考える．

splog の定義をこのようにした理由は，Table 2,3 に示す実際の splog がこのようにして生成されているからであり，この定義に従って splog とみなすことができるからである．

3.2 作成方法

本稿で提案する splogger 検出法を評価するために，splog ベンチマークデータを作成した．データは，2009 年 1 月 15 日 から 2008 年 1 月 25 日 までの間，日本国内の大手ブログ CSP(Content Service Provider) からユーザをサンプリングした．各社が配信している新着ブログエントリの RSS フィードを常時監視し，新着ブログエントリを投稿したユーザを列挙し，その中からサンプリング率　0.0005 でユーザのランダムサンプリングを行った．

次に，収集したブログ（web ページを）を人手で目視し splogger/blogger の判定を行い，計 576 件のラベル付きリストを作成した．Table 2,3 に示すように splog 分類に基づき，各種別の統計をとった．これの統計を Table 1，Table 4，Fig 1 に示す．以前までの調査に比べ，movie(Table3) に分類される splog が多くなってきたため，これを content snatch から分割した．

なお，複製文字列検知に基づいた Splog フィルタリング手法 [6] で述べたとおり，Table3 の 1 種の方法だけで生成されている splog ばかりではなく，複数の生成手法を組み合わせている (Table2　combine　) 場合もある．今回のケースでは，product

splogger	164	28.47 %
blogger	412	71.52 %
total	576	

Table 1　splogger/blogger 構成比

content snatch	他のコンテンツをコピーしてくる Table3 のような 様々な亜種が存在する
search results	ある単語 (X) の検索結果コンテンツ の中身をコピーする. 単語 X は，自動投稿ツールが 持っている辞書の単語を使う場合と， 最近話題のキーワードを API で取得する場合とがある. 単語 X は複数の単語の場合もある
word salad	単語を無数に並べる. 自動投稿ツールが内部的に持って いる辞書の単語を使う場合と 最近話題のキーワードを API で 取得する場合と， その両方を使う場合とがある.
template decorator	コンテンツに テンプレート文字列を付加する. 例えば「お早うございます」 「今日は……でした」 「……をしています」 等
combine	Table 2 ,Table3 の生成手法を 2 種類以上組み合わせる

Table 2　splog 分類表 1

induction + word salad , template decorator + news update , template decorator + word salad , などの複合手法を確認した.

3.3　ブログエントリの本文抽出

ブログエントリ中の広告などを除いて本文のみを抽出するために，html のタグを用いた．文献 [5] と同様に，それぞれの CSP の本文タグを調べ，その中のテキストだけを取得することにより十分な信頼性で本文を抽出することができる．同様の方法でブログの説明文も取得できる.

news update	web ニュースの記事本文を すべてコピーして生成する.
mail magazine	メールマガジンの記事本文を すべてコピーして生成する.
dictionary	wikipedia などの百科事典コンテンツを そのままコピーして生成 する.
QA	yahoo 知恵袋, はてな人力検索, などの 質問文 (場合によっては回答文も) をそのままコピーする.
product induction	EC サイトの商品販売ページと 「機能レベルで」ほとんど同じであり そこから商品を直接購入もできる.
RSS	特定の RSS(Rich/RDF Site Summary , Really Simple Syndication) の内容をそのままコピーする 誰でもアクセス可能な RSS や splogger が設定した RSS の場合もある
movie	Youtube など動画サイトからの 動画参照しかコンテンツがない

Table 3　splog 分類表 2

dictionary	2
movie	5
news update	10
product induction	40
search results	57
content snatch (other)	6
template decorator + word salad	5
word salad	1
product induction + word salad	33
template decorator + news update	5

Table 4　splog 種別統計

ブログの説明文は， Fig 2，Fig 3 に示すようにブログのタイトルとともにページ上部に表示される場合が非常に多く，ブログタイトルと同様に，全ての個別ページに共通で表示される.

実際はこの説明文に該当するフォーマットが存在しないブログも少数ではあるが存在する．そのような場合は html ソース中の meta タグ中の，description に該当する部分の内容を取得する．説明文がある場合は同じ内容がここに記述され，説明文がない場合も description にはブログの説明文として適切な内容が記述されている場合が非常に多いためである.

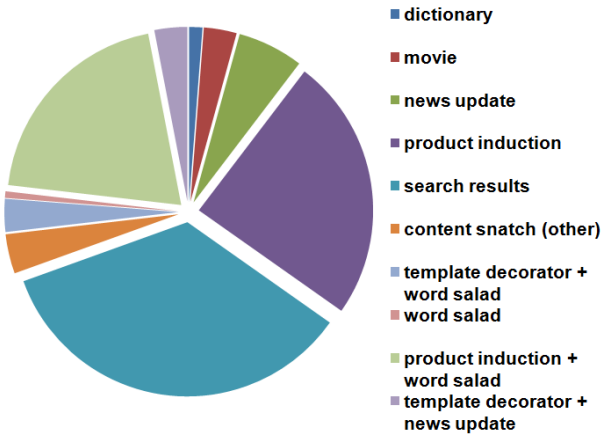

Fig. 1　splog 種別統計

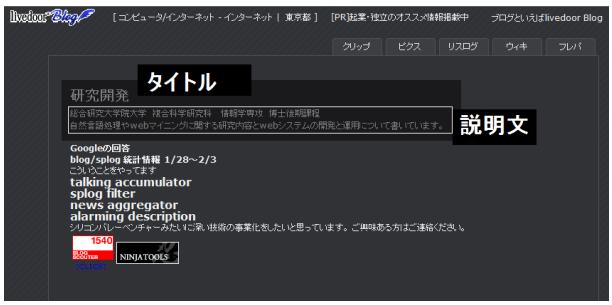

Fig. 2　blog 説明文 1

3.4　BlogHeader

こうしてブログのタイトルと説明文をブログの特徴量としたコーパスを作成した．

3.5　BlogContent

タイトルが本当に有効な特徴量となるのか，比較のため BlogHeader と同じく，576 件のブログそれぞれの，2009 年 2 月 1 日 時点での 各 blogger の最新のブログエントリのコンテンツ本文を特徴量としたコーパスを作成した．splogger/blogger のラベルは，BlogHeader と同じである．

blogger が splog の定義を満たすコンテンツを作成するケースは多々あるが，今回のコーパス（BlogContent）中では splogger/blogger と splog/blog のラベルが一致しないケースはなかった．

4　splogger 判別手法

本稿では splogger の特定を行うための特徴量の有効性と，特徴量の傾向を示すため単純なモデルを用いたい．このため，最もシンプルなナイーブベイズ分類器による判定を選択した．

Fig. 3　blog 説明文 2

4.1　ナイーブベイズ分類器

ナイーブベイズ分類器とは，独立性仮定と共にベイズの定理を適用することに基づいた単純な確率的分類器である．クラス $c_i(1 \leq i \leq t)$ の事前確率 $P(c_i)$ と素性 $x = (x_1, x_2, ..., x_n)$ の条件付き確率 $P(x|c_i)$ を最大化する c_i を求める問題として定式化される．

$$c_i = argmax_{c_i} P(c_i) P(x_1, x_2, ..., x_n | c_i) \tag{1}$$

全ての単語 x_j が他の単語と完全に独立に現れると仮定することでこの式を単純化できる．

$$c_i = argmax_{c_i} P(c_i) \prod_j P(x_j | c_i) \tag{2}$$

ここで，

$$P(c_i) = \frac{\text{クラス } c_i \text{ のブログ数}}{\text{全ブログ数}} \tag{3}$$

$$P(x_j|c_i) = \frac{\text{クラス } c_i \text{ での数 } x_j \text{ の出現頻度}}{\text{クラス } c_i \text{ のブログ数}} \tag{4}$$

今回は，$c_1 = splogger$, $c_2 = blogger$ の 2 クラスのみである．

なお，あらかじめすべての単語の出現頻度に 0.5 を加算することで，ゼロ頻度問題には対応した．

今回は素性 $x = (x_1, x_2, ..., x_n)$ として，各ブログの本文 (BlogContent) または説明文 (BlogHeader) の文章を形態素解析した単語をすべて用いる．

評価指標	
precision	0.827
recall	0.63
F	0.715
splog 種別 recall	
dictionary	0
movie	0.6
news update	0.4
product induction	0.671
search results	0.632
content snatch (other)	0.833
template decorator + word salad	0.8
word salad	0
product induction + word salad	0.671
template decorator + news update	0.4

Table 5　BlogHeader の精度

評価指標	
precision	0.428
recall	0.979
F	0.595
splog 種別 recall	
dictionary	1
movie	1
news update	1
product induction	0.984
search results	0.982
content snatch (other)	1
template decorator + word salad	1
word salad	1
product induction + word salad	0.984
template decorator + news update	1

Table 6　BlogContent の精度

5　評価実験

5.1　実験の概要

3 章で構築したラベル付きコーパス BlogHeader, BlogContent に対して，4 章で提案した 判別手法を用いて各 blogger を判定した結果を調べた．

今回は交差検定 (cross-validation) によって splogger 判定の評価を行う．ラベル付きデータを 10 分割し，その内 9 個を学習用として頻度情報とし，残りの一つに対する正解判定を行う．これを 10 回分繰り返し，その結果の平均値をとる．

5.2　評価指標

splogger 判定の性能をはかる評価指標を以下に示す．

$$recall \equiv \frac{\text{システム出力と正解ラベルが splog であると一致する数}}{\text{データ中のブログエントリ (splog) 数}} \tag{5}$$

$$precision \equiv \frac{\text{システム出力と正解ラベルが splog であると一致する数}}{\text{システムがブログエントリ (splog) であると出力した数}} \tag{6}$$

$$F\ 値 \equiv \frac{2\,recall\,precision}{recall + precision} \tag{7}$$

5.3　結果と考察

BlogContent では 1.0 に近い recall を実現しているが，precision は 0.5 以下であり，半分以上当たらないという結果である．これに対して BlogHeader では 0.8 以上の precision を実現できており，本文と説明文との，特徴量としての有効性の差は示せたと思う．本文の内容からでは，通常の文書分類などの方法では splog 判定できない場合が非常に多いのである．

splog 種別に有効性の差が生じるかであるが，Table 5 , 6 の結果からは必ずしも一般的な傾向は見いだせない．実際は各 splog の種類の他に説明文にも複数の分類種別があり，splog の種類が同じであるからと言って，説明文の傾向も同じというわけではない．説明文にほとんど何も入力しない splogger も存在する．

具体的には splogger のブログ説明文にはどのような説明文が使われるかというと,「○×のご紹介」「○×の最新情報」「○×の口コミ情報」などである．注意していただきたいのは，splogger 特有の単語とは「○×」の部分ではない,「○×」には人名，商品名など固有名詞が入る場合が多いが，splog であることと「○×」という単語が入っていることはほとんど関係ない「ご紹介」「最新情報」「口コミ情報」などの方こそが，splogger であると判断

する根拠としては強いのである．

6 おわりに

本稿ではブログタイトル，ブログ説明文などメタデータに注目した splog フィルタリング手法を提案した．提案手法は，ブログの本文を一切考慮せず，ブログタイトルと説明文のみから splog 判定を行う．使用する特徴量となる文字列の量も少なく，判定手法も，最もシンプルなベイズフィルタであり，非常に軽量な手法である．また，提案手法はブログエントリ単位で一つ一つ判定するのではなく，blogger 単位での判別を行うので，判定効率は非常に良い．

しかし一方で，提案手法には正解データの構築が必要であり，提案手法を実データに適用するためには，何らかの方法による（半）自動化が必要である．また，precision，recall など判定精度の点ではまだまだ改善が望まれる．

参考文献

1) 奥村学, blog マイニング： インターネット上のトレンド , 意見分析を目指して , 人工知能学会誌", volume 21 , no 4 , pp 424-429, (2006).

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),(2007).

3) P. Kolari and A. Java and T. Finin, Characterizing the Splogosphere, Proc. of 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics,(2006).

4) T. Takeda and T. Takasu, A Spam Blog Filitering Method Based on Text Copy Detection , The First IEEE International Conference on the Applications of Digital Information and Web Technologie, 543-548 (2008).

5) T. Takeda and T. Takasu, UpdateNews: a news clustering and summarization system using efficient text processing, International Conference on Digital Libraries (JCDL 2007), pp 438-439, (2007).

6) 竹田隆治, 高須淳宏, 複製文字列検知に基づいた Splog フィルタリング手法 , 情報処理学会論文誌：データベース (TOD41-19) (2009).

7) Y. R. Lin and Wen-Yen Chen and Xiaolin Shi and Richard Sia and Xiaodan Song and Y. Chi, and K. Hino and H. Sundaram and J. Tatemura and B. Tseng , The splog detection task and a solution based on temporal and link properties , Proc. of 15th Text REtrieval Conference (TREC'06) , 2006.