2E2-3in

機械学習・アフィリエイトID・HTML構造の類似性の併用によるスプログ検出

Detecting Splogs by Integrating Machine Learning / Affiliate IDs / Similarity of HTML Structures

森尻 惇宜史[*1] 片山 太一[*2] 宇津呂 武仁[*1] 河田 容英[*3] 福原 知宏[*4]
Akihito Morijiri Taichi Katayama Takehito UTSURO Yasuhide Kawada Tomohiro FUKUHARA

[*1]筑波大学大学院システム情報工学研究科 知能機能システム専攻
Graduate School of Systems and Information Engineering,University of Tsukuba

[*2]日本電信電話株式会社 NTT サイバースペース研究所 [*3](株) ナビックス
NTT Cyber Space Laboratories, NTT Corporation Navix Co., Ltd.

[*4]独立行政法人 産業技術総合研究所 サービス工学研究センター
Center for Service Research, National Institute of Advanced Industrial Science and Technology

Spam blogs or splogs are blogs hosting spam posts, created using machine generated or hijacked content for the sole purpose of hosting advertisements or raising the number of in-links of target sites. It has been shown that splogs can be detected based on similarity of HTML structures and affiliate IDs. The similarity of HTML structures of splogs is effective in splog detection, and the identity of affiliate IDs extracted from splogs can identify spammers much more directly than similarity of HTML structures, although it is not easy to achieve high coverage in extracting affiliate IDs. The coverage of the intersection of the two clues, similarity of HTML structures and affiliate IDs, is relatively low, and it is necessary to apply them in a complementary strategy. This paper studies how to detect splogs which cannot be detected based on either similarity of HTML structures nor affiliate IDs. We apply SVMs to this task and show that splogs of above type can be detected with high precision.

1. はじめに

ブログには個人の意見情報が記されており，市場の動向を推測するための手掛かりや製品についての意見調査をする上で有益であるとして，近年注目を集めている．そのため，従来からあるインデクシングのみを行う検索エンジンとは異なる，ブログ特有の情報検索サービスが出現している．具体的には，ブログ解析サービスとして，Technorati，BlogPulse [Glance 04]，kizasi.jp などが存在する．多言語ブログサービスとしては，Globe of Blogs が言語横断ブログ記事検索機能を提供している．また Best Blogs in Asia Directory がアジア言語ブログの検索機能を提供している．一方で，ブログのウェブコンテンツの作成と配信は非常に容易になっており，そのことが引き金となって，アフィリエイト収入を得ることを目的とするスパムブログ (以下，スプログ) が急増している [Gyöngyi 05, Kolari 06b, Macdonald 06, Kolari 07]．スプログにおいては，通常，広告主への誘導または対象サイトの被リンク数を増加する目的のもとで，機械的な文書作成や他サイトの引用という手段を用いて自動的に記事を生成し，大量のリンクを有するブログを機械的に自動生成する．[Kolari 06b] は英語ブログにおいて，約 88%のブログサイトがスプログであり，それは全ブログポストの 75%を占めると報告している．このことから，[Lin 07] に述べられているように，スプログは情報検索品質の低下やネットワークと格納資源の多大な浪費などといった問題を起こす要因となる．そのため，近年，スプログの分析や検出を目的とした研究が進められている．いくつかの既存研究 [Kolari 06b, Macdonald 06, Kolari 07] はスプログの重要な特性を報告している．[Macdonald 06] では，TREC Blog06 データコレクションを用いて，スプログのピング時系列特性，入力度数/出力度数の分布特性，典型的な単語群を分析している．また，[Kolari 06b, Kolari 07] は，*BlogPulse* データセットを用いたスプログ分析の結果を報告している．一方，[石田 08, Kolari 06a, Mishne 05, Lin 07] 等においては，言語情報，リンク情報，HTML タグ情報，時間情報といった多様な特性を手がかりとしてスプログを検出する技術を提案している．

上記の既存研究とは異なり，HTML 構造の類似性やアフィリエイト ID を手がかりとして，スプログの検出を行なう研究もある．スプログにおいては，一人の作成者が複数のスプログを機械的に生成していると考えられる．そこで，[片山 10b] では，同一の作成者によって作成されたスプログの組において，HTML 構造が類似している場合があり，その特性を利用したスプログの検出を行なうことができることを報告している．これは，図 1 における領域「1」,「3」を適用範囲とする手法になる．また，[石井 10] においては，アフィリエイト ID をスプログから自動抽出し，複数のブログサイトに含まれるアフィリエイト ID に着目して，スプログを収集，分析する手法を提案している．これは，図 1 における領域「2」,「3」を適用範囲とする手法になる．これらの 2 つの手がかりについて，[片山 10a, Katayama 11] では，それぞれの手がかりの適用範囲の違いを示し，これらの 2 つの手がかりは相補的に用いる必要性があることを報告している．以上をふまえて，本論文では，HTML 構造の類似性，アフィリエイト ID のいずれによっても検出できないスプログ (図 1 における領域「4」に対応する) に対して，機械学習の一つである Support Vector Machines (SVMs) [Vapnik 98] を用いることで，高適合率でスプログが検出できることを示す．

連絡先: 森尻惇宜史，筑波大学大学院システム情報工学研究科，
〒 305-8573 茨城県つくば市天王台 1-1-1，029-853-5427

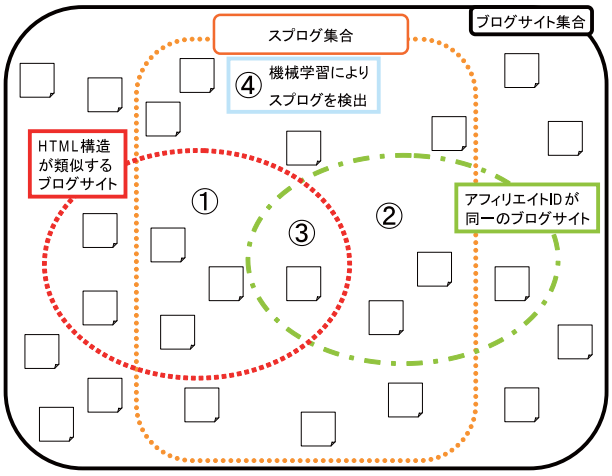


図 1: スプログ検出における各手法の適用範囲

表 1: 10 以上のスプログに含まれるアフィリエイト ID 数およびそれらのアフィリエイト ID を含むスプログの総数

ブログホスト H	アフィリエイト ID 数	$\left\lvert \bigcup_x SP_{af}(H,x) \right\rvert$
S 社	72	1,101
F 社	56	953

図 2: アフィリエイト ID を含むアフィリエイトリンクの例

2. アフィリエイト ID を用いたスプログの分析

図 2 にアフィリエイト ID の抽出例を示す．アフィリエイトリンクには，そのアフィリエイトリンクを生成したアフィリエイタのアフィリエイト ID や，広告主の ID，商品 ID などが含まれており，我々はその中からアフィリエイト ID の抽出を行った．本論文では，特に，[石井 10] にしたがい，ASP(アフィリエイト・サービス・プロバイダ) のうち実際にアフィリエイト ID の抽出が可能な 10 社[*1]を対象としてアフィリエイト ID の抽出を行った．

日本語ブログ収集にあたり，中国語，日本語，韓国語，英語のブログ記事の収集を行う KANSHIN システム [福原 07] を利用する．このシステムでは，各言語のブログサイトのリストを参照し，ブログサイトの提供する RSS フィードファイルと Atom フィードファイルを取得し，記事をデータベースに蓄積している．このシステムに蓄積されたサイトから，分析対象とした S 社，F 社のブログホスト会社 2 社について，合わせて約 11 万ブログサイトを収集し，S 社の 48,183 サイトのうち 14,352 サイト (約 30%) から，F 社の 60,977 サイトのうち 6,231 サイト (約 10%) から，それぞれアフィリエイト ID が抽出された．

ここで，スパマーは，ASP から得られる報酬を少しでも増やすために，一つのアフィリエイト ID に対して複数のスプログを作成していると考えられる．そのため，複数のブログサイトに出現するアフィリエイト ID は，スパマーがスプログにおいて使用しているアフィリエイト ID である可能性が高くなると考えられる．あるアフィリエイト ID が，スパマーがスプログにおいて使用しているアフィリエイト ID であると推定できれば，そのアフィリエイト ID が出現する全てのブログサイトはスプログと判断することができる．以上の考え方に基づき，[石井 10] においては，複数のブログサイトに出現するアフィリエイト ID を分析し，実際にスパマーがスプログにおいて使用しているアフィリエイト ID であると推定できる割合を報告している．本論文においても，[石井 10] の分析結果をふまえて，同一のアフィリエイト ID を含むブログサイト数が多いほど，スプログを自動生成している可能性が高いと考えて，ASP10 社のうちいずれかのアフィリエイト ID が抽出された 20,583 サイト (S 社，F 社の合計) を対象として，10 以上のブログサイトから抽出されたアフィリエイト ID を分析対象とした．

その結果，129 個のアフィリエイト ID が抽出され，これらのアフィリエイト ID を含むブログサイト数は 2,472 となった．このうち，スパマーがスプログにおいて使用しているアフィリエイト ID であると推定できたアフィリエイト ID は 121 個であり，これらのアフィリエイト ID を含むスプログ数は 2,054 であった (アフィリエイト ID のスパム率は 93.8%，ブログサイトのスプログ率は 83.3%)[*2]．ここで，この 121 個のアフィリエイト ID の各々を x として，ブログホスト H においてアフィリエイト ID x を含むスプログの集合を $SP_{af}(H,x)$ と定義する．また，各ブログホストに出現したアフィリエイト ID 数，および，いずれかのアフィリエイト ID を含むスプログの総数を表 1 に示す．

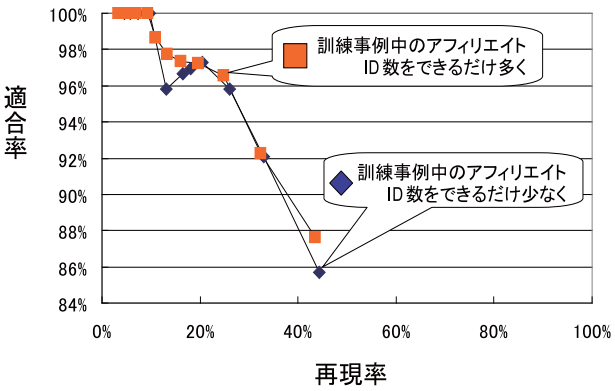
3. スプログ検出および信頼度尺度

SVM による機械学習を行うためのツールとして，TinySVM (`http://chasen.org/~taku/software/TinySVM/`) を用いた．スプログ検出のための素性としては，[森尻 11] のものを用いる．カーネル関数としては，線形および 2 次多項式を比較し，2 次多項式の方が性能が良かったため，4. においては，2 次多項式カーネルを用いた場合の結果を示す．また，全ての素性に値がないものは訓練データから除外する．

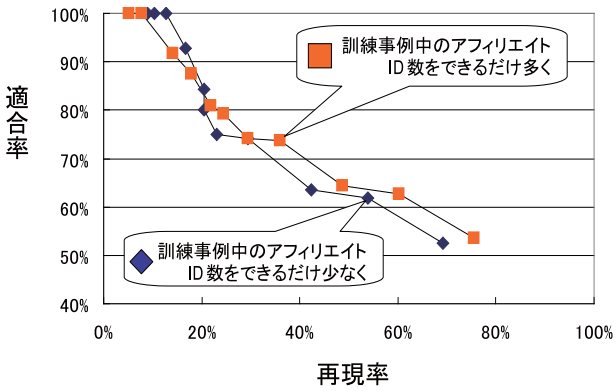
また，SVM による機械学習での信頼度尺度として，分離平面から各評価事例への距離を用いた [Tong 00]．具体的には，スプログとして判定される事例に対する分離平面からの距離の下限 LBD_s および，非スプログとして判定される事例に対する分離平面からの距離の下限 LBD_{ab} をそれぞれ設定する．

*1 Am 社，At 社，D 社，Gl 社，I 社，Lk 社，R 社，St 社，Tr 社，V 社．

*2 複数のブログサイトに含まれるアフィリエイト ID のうち，スパマーがスプログにおいて使用しているアフィリエイト ID とみなすことができないものを大別すると，自動アフィリエイト作成ツールの作成者のアフィリエイト ID の場合と，ブログホスト会社のアフィリエイト ID の場合とに分けられる．両者とも，スパマーがスプログにおいて使用しているアフィリエイト ID と比較して，アフィリエイト ID が出現するブログサイト数が相対的に多いため，アフィリエイト ID のスパム率よりもブログサイトのスプログ率の方が低くなった．

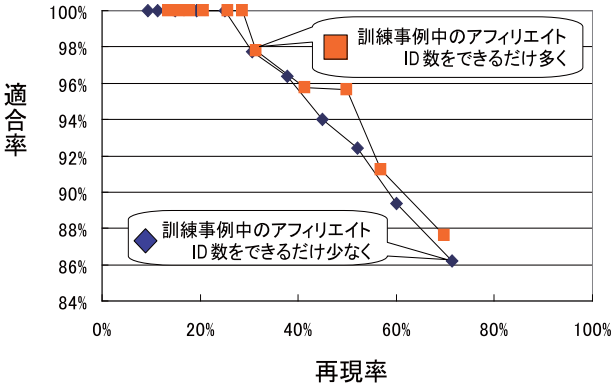


(a) スプログ検出性能

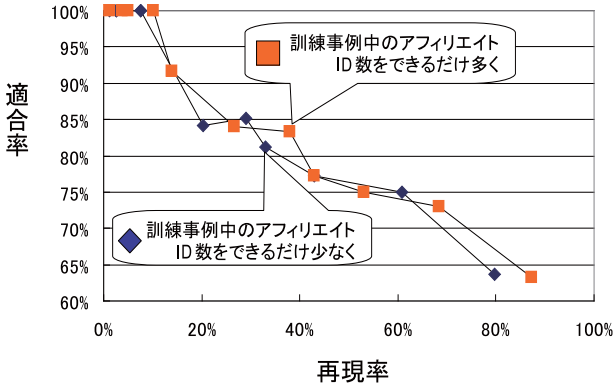


(b) 非スプログ検出性能

図 3: スプログ・非スプログ検出性能 (S 社)



(a) スプログ検出性能



(b) 非スプログ検出性能

図 4: スプログ・非スプログ検出性能 (F 社)

4. 評価

本節では，複数のブログサイトに含まれるアフィリエイト ID に着目して収集したスプログ (図 1 における領域「2」および「3」のスプログ)，および，無作為に収集した非スプログを訓練事例として SVM によりスプログ検出器の訓練を行い，HTML 構造の類似性，アフィリエイト ID のいずれによっても検出できないスプログ (図 1 における領域「4」のスプログ) を検出する評価実験を行った結果について述べる．

4.1 訓練事例

2. 節で収集したブログサイトが，S 社は 48,183 サイト，F 社は 60,977 サイトあり，これらのうちで，10 以上のブログサイトから抽出されたアフィリエイト ID を含むスプログ数は，表 1 に示すように，S 社が 1,101 サイト，F 社が 953 サイトであった．

機械学習においては一定数以上の訓練事例が必要である．ここで，アフィリエイト ID を手がかりとして一定数の訓練事例を収集する場合に，以下の二通りの考え方がある．

- 訓練事例に含まれるアフィリエイト ID の数をできるだけ多くなるようにする
- 訓練事例に含まれるアフィリエイト ID の数をできるだけ少なくなるようにする

本論文では，この二通りの考え方に基づいた訓練事例の収集をそれぞれ行い，検出性能の比較も行う．

訓練事例に含まれるアフィリエイト ID の数ができるだけ多くなるようにした場合には，訓練事例収集のコストは最大になる．この場合，訓練事例の収集を行う際には，全てのアフィリエイト ID を選択し，S 社については 72 個のアフィリエイト ID を，F 社については 56 個のアフィリエイト ID を，それぞれ選択する．その後，全てのアフィリエイト ID について，それらを含むブログサイトを各アフィリエイト ID について同数程度ずつ選択し，500 サイト (各ホストあたり) を SVM におけるスプログの訓練事例とする．

一方，訓練事例に含まれるアフィリエイト ID の数ができるだけ少なくなるようにした場合には，訓練事例収集のコストは最小になる．この場合，訓練事例の収集を行う際には，より多くのブログサイトに含まれるアフィリエイト ID から順に選択し，S 社については 13 個のアフィリエイト ID を，F 社については 12 個のアフィリエイト ID を，それぞれ選択する．その後，より多くのブログサイトに含まれるアフィリエイト ID を含むブログサイトから順に，これらのアフィリエイト ID を含むブログサイトを 500 サイト (各ホストあたり) を選択して，これを SVM におけるスプログの訓練事例とする．

また，あらかじめ人手で判定した非スプログを 500 サイト用意し[*3]，これを非スプログの訓練事例とした．

*3 非スプログの収集の際に，本論文では無作為に収集を行なった．しかし，非スプログの作成者は一つしかブログを作成しないという考え方に基づいて，一つのブログサイトにのみ含まれるアフィリエイト ID を手がかりとして，非スプログの訓練事例をより簡単に収

4.2 評価事例

2. 節で収集したブログサイトは，S 社においては 48,183 サイト，F 社においては 60,977 サイトであった．これらのうち，10 以上のブログサイトから抽出されたアフィリエイト ID を含まず，かつ，10 以上のブログサイトから抽出されたアフィリエイト ID を含むスプログと HTML 構造が類似しない (MinDF > 0.15)*4 ブログサイトは，S 社は 47,029 サイト，F 社は 59,982 サイトとなった．このブログサイト集合を評価事例の候補集合とすることで，図 1 における領域「4」のスプログを対象として検出を行う．

この候補集合のうち，分離平面からの距離が分散されるようにスプログ側，非スプログ側の両方から評価事例を選択し，S 社については 283 サイト，F 社については 268 サイトに対して，人手でスプログ/非スプログの判定を付与した．

4.3 評価手順および評価尺度

スプログとして判定される事例に対する分離平面からの距離の下限 LBD_s について，分離平面からの距離が LBD_s 以上となる評価事例に対して，スプログと判定した場合の再現率，適合率を測定する．そして，LBD_s を変化させた場合の再現率，適合率の推移を，S 社については図 3(a) に，F 社については図 4(a) にプロットした．

また，非スプログについても同様に，非スプログとして判定される事例に対する分離平面からの距離の下限 LBD_{ab} について，分離平面からの距離が LBD_{ab} 以上となる評価事例に対して，非スプログと判定した場合の再現率，適合率を測定する．そして，LBD_{ab} を変化させた場合の再現率，適合率の推移を S 社については図 3(b) に，F 社については図 4(b) にプロットした．

また，図 3，および，図 4 においては，訓練事例中のアフィリエイト ID 数をできるだけ多くした場合については，「訓練事例中のアフィリエイト ID 数をできるだけ多く」としてプロットし，訓練事例中のアフィリエイト ID 数をできるだけ少なくした場合については，「訓練事例中のアフィリエイト ID 数をできるだけ少なく」としてプロットした．

4.4 評価結果

図 3，および，図 4 に示すように，訓練事例中のアフィリエイト ID 数をできるだけ多くした場合と，できるだけ少なくした場合を比較した時，再現率，適合率の推移はほぼ同等となった．これにより，アフィリエイト ID を用いて訓練事例を収集する際には，収集のコストを抑えるために訓練事例に含まれるアフィリエイト ID を少なくしても，検出の再現率や適合率に与える影響は少ないことが分かった．

また，図 3(a)，および，図 4(a) に示すように，スプログの検出において，S 社においては再現率約 35%以下の範囲で，F 社においては再現率約 60%以下の範囲で，90%を超える高い適合率となった．一方，非スプログの検出においては，図 3(b)，および，図 4(b) に示すように，S 社においては再現率約 17%以下の範囲で，F 社においては再現率約 15%以下の範囲で，90%を超える高い適合率となった．

4.2 節で示したように，今回の評価事例は図 1 における領域「4」を対象としている．これにより，HTML 構造の類似性，アフィリエイト ID のいずれによっても検出できないスプログに対して，高い適合率での検出を実現することができた．しかも，その適合率は，HTML 構造の類似性 [片山 10b] やアフィリエイト ID [石井 10] の手がかりを単独で用いてスプログの検出を行なう場合とほぼ同等の高い適合率となった．

集する手法も考えられる．

*4 MinDF の定義の詳細は [片山 10b] を参照のこと．

5. おわりに

アフィリエイト収入を得ることを目的とするスプログの検出タスクにおいて，これまで，HTML 構造の類似性やアフィリエイト ID を用いることにより，一定の範囲のスプログが検出できることが知られていた．これらの手法は単独で用いた場合の適用範囲が十分ではなく，両者の手がかりを併用する必要があった．これに対して，本論文では，HTML 構造の類似性，アフィリエイト ID のいずれによっても検出できないスプログに対して，SVM を適用することにより，高適合率の検出が可能であることを示した．

参考文献

[福原 07] 福原 知宏, 宇津呂 武仁, 中川 裕志, 武田 英明：複数の言語で記述されたブログ記事を対象とした言語横断型関心システム, 第 21 回人工知能学会全国大会論文集 (2007)

[Glance 04] Glance, N., Hurst, M., and Tomokiyo, T.: BlogPulse: Automated Trend Discovery for Weblogs, in *Proc. Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics* (2004)

[Gyöngyi 05] Gyöngyi, Z. and Garcia-Molina, H.: Web Spam Taxonomy, in *Proc. 1st AIRWeb*, pp. 39–47 (2005)

[石田 08] 石田 和成：スパムブログの推定と抽出, 日本データベース学会 Letters, Vol. 6, No. 4, pp. 37–40 (2008)

[石井 10] 石井 聡一, 福原 知宏, 増田 英孝, 中川 裕志：アフィリエイト ID を用いたスパムブログの分析, WebDB2010 論文集 (2010)

[片山 10a] 片山 太一, 森尻 惇宜史, 石井 聡一, 宇津呂 武仁, 河田 容英, 福原 知宏：HTML 構造の類似性およびアフィリエイトを用いたスプログの分析, WebDB2010 論文集 (2010)

[片山 10b] 片山 太一, 芳中 隆幸, 宇津呂 武仁, 河田 容英, 福原 知宏：HTML 構造を利用した類似スパムブログの収集, 第 2 回 DEIM フォーラム論文集 (2010)

[Katayama 11] Katayama, T., Morijiri, A., Ishii, S., Utsuro, T., Kawada, Y., and Fukuhara, T.: Comparing Similarity of HTML Structures and Affiliate IDs in Splog Analysis, in Xu, J., et al. eds., *Proc. 16th DASFAA, Inter. Workshops: SNSMW*, Vol. 6637 of *LNCS*, pp. 378–389, Springer (2011)

[Kolari 06a] Kolari, P., Finin, T., and Joshi, A.: SVMs for the Blogosphere: Blog Identification and Splog Detection, in *Proc. 2006 AAAI Spring Symp. Computational Approaches to Analyzing Weblogs*, pp. 92–99 (2006)

[Kolari 06b] Kolari, P., Joshi, A., and Finin, T.: Characterizing the Splogosphere, in *Proc. 3rd Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics* (2006)

[Kolari 07] Kolari, P., Finin, T., and Joshi, A.: Spam in Blogs and Social Media, in *Tutorial at ICWSM* (2007)

[Lin 07] Lin, Y.-R., Sundaram, H., Chi, Y., Tatemura, J., and Tseng, B. L.: Splog Detection using Self-similarity Analysis on Blog Temporal Dynamics, in *Proc. 3rd AIRWeb*, pp. 1–8 (2007)

[Macdonald 06] Macdonald, C. and Ounis, I.: The TREC Blogs06 Collection : Creating and Analysing a Blog Test Collection, Technical Report TR-2006-224, University of Glasgow, Department of Computing Science (2006)

[Mishne 05] Mishne, G., Carmel, D., and Lempel, R.: Blocking Blog Spam with Language Model Disagreement, in *Proc. 1st AIRWeb* (2005)

[森尻 11] 森尻 惇宜史, 片山 太一, 石井 聡一, 宇津呂 武仁, 河田 容英, 福原 知宏：スプログの収集における HTML 構造の類似性およびアフィリエイトの分析, 第 3 回 DEIM フォーラム論文集 (2011)

[Tong 00] Tong, S. and Koller, D.: Support Vector Machine Active Learning with Applications to Text Classification, in *Proc. 17th ICML*, pp. 999–1006 (2000)

[Vapnik 98] Vapnik, V. N.: *Statistical Learning Theory*, Wiley-Interscience (1998)