HTML 構造の類似性を利用した大規模スパムブログ収集*

片山 太一† 芳中 隆幸‡ 宇津呂 武仁† 河田 容英§ 福原 知宏¶
筑波大学大学院 システム情報工学研究科† ，東京電機大学大学院 工学研究科‡
(株) ナビックス§ ，東京大学 人工物工学研究センター¶ ，

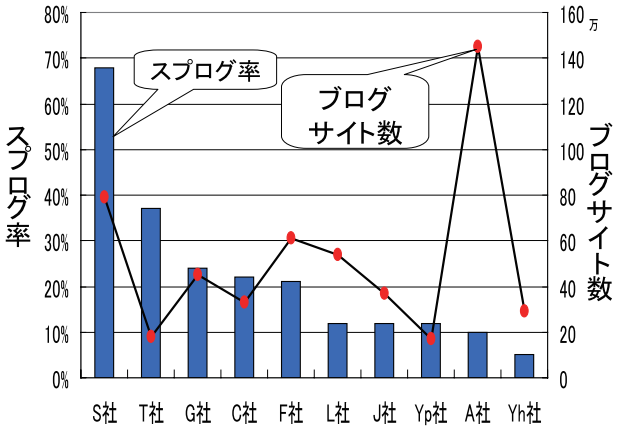


図 1: ホストごとのブログサイト数およびスプログ率

1 はじめに

本研究では，ブログにおいてアフィリエイト収入を得ることを目的とするスプログ (スパムブログ) [3] のうち，特に，同一の作成者が自動的に大量生成したと推測されるスプログの検出において，HTML 構造の類似性が効果的であることを示す．具体的には，ブログの HTML ファイルにおける DOM ツリーから，コンテンツの最小単位に相当するブロックを抽出し，複数のスプログの間でブロック構造の類似性を測定する．その結果，同一ブログホストにおけるスプログのうち，同一のスプログ作成者が自動的に大量生成したと推測されるスプログ [4] 同士では，ブロック構造が類似する傾向があることを示す．さらに，この特性を利用して，日本語ブログ空間中におけるスプログを大規模に収集する方式を提案する．

2 HTML 構造を用いたスプログ検出

文献 [1] で提案された手法により，HTML ファイルから DOM 系列を抽出し，その差分の割合を計算する．

表 1: ホストごとの教師なしスプログ検出可能性の分析

HTML 構造のみを用いたスプログ検出: 高適合率で可能		HTML 構造のみを用いたスプログ検出: 高適合率では困難	
$AvMin DF_{10}$ の上限値が 0.15〜0.3	$AvMinDF_{10}$ の上限値が 0.05〜0.15，ブログのテンプレートに制限がある	各ホスト 500 ブログサイト中，同一作成者によって大量生成されたスプログのサンプル数が少ない	非スプログ同士の HTML 構造が似る
C 社，S 社，T 社	A 社，G 社，Yp 社	F 社，J 社	L 社，Yh 社

表 2: ホストごとの教師なしスプログ検出性能

ブログホスト会社	C 社	S 社	T 社	A 社	G 社	Yp 社
適合率 (%)	94	100	95	100	100	83
再現率 (%)	32	23	74	22	65	17

文献 [2] で作成した主要ブログホスト会社 10 社の日本語スプログ・非スプログデータセット中の各 500 ブログサイトに対して[1]，文献 [1] で導入した $AvMinDF_k(s,T)$ (ただし，$k=10$) に対して上限値を設け，

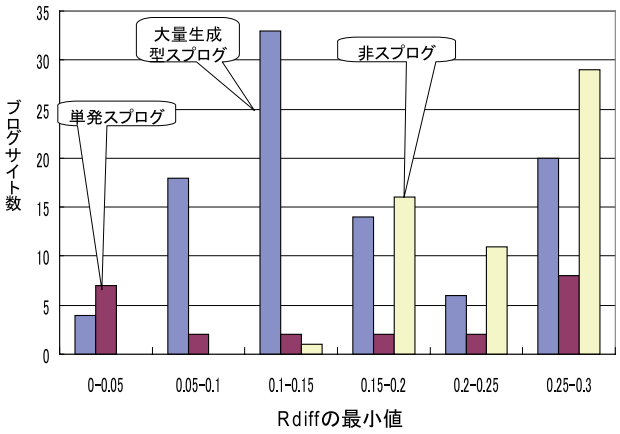
> $AvMinDF_{10}(s,T)$ の値が上限以下となるブログサイトはスプログである

という規則により，教師なしスプログ検出を行った．10 社のブログホスト会社は，表 1 に示すように 4 種類の傾向に分かれた．10 社のうち，6 社はスプログ検出が高適合率で可能であるが，4 社は困難であるという結果になった．スプログ検出が高適合率で可能な 6 社は $AvMinDF_{10}(s,T)$ の閾値の違いにより 2 種類に分けられる．表 2 に高適合率でスプログ検出が可能であった 6 社の適合率と再現率を示す．高適合率でのスプログ検出が不可能な 4 社について，大きく 2 種類に分けられる．F 社および J 社については，各ホスト 500 ブログサイト中，同一作成者によって大量生成されたスプログのサンプル数が少なく，$Rdiff$ の値が最小の 10 サイトの中に，

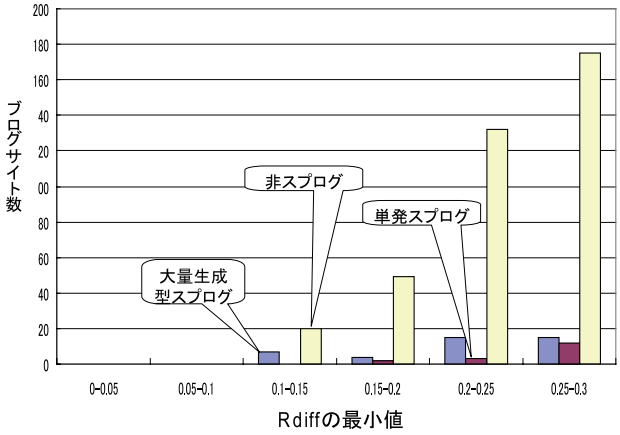
*Large-scale Automatic Collection of Splogs based on Similarities of HTML Structures

†Taichi Katayama, Takahito Utsuro, Graduate School of Systems and Information Engineering, University of Tsukuba,
‡Takayuki Yoshinaka Graduate School of Engineering, Tokyo Denki University,
§Yasuhide Kawada, Navix Co., Ltd.
¶Tomohiro Fukuhara, Research into Artifacts, Center for Engineering, University of Tokyo,

[1]図 1 に，ホストごとのブログサイト数，ラベル付けを行ったブログのスプログ率を示す．



(a) スプログとの $Rdiff$ の最小値



(b) 非スプログとの $Rdiff$ の最小値

図 2: 52 万ブログサイトを対象とする $Rdiff$ の最小値の分布 (F 社)

同一作成者以外によって作成されたスプログや非スプログが混入していた．しかし，$Rdiff$ の値が最小の 10 サイトの中にも，同一作成者によるスプログで，しかも，DOM 系列差分の割合が小さいものが存在することも確認できた．したがって，$Rdiff$ を計算する対象とするブログサイト集合を大規模化することにより，スプログ検出の適合率を改善できると期待できる．

一方，L 社および Yh 社については，非スプログ同士の HTML 構造が類似し，高適合率でのスプログ検出は困難であった．これらのブログホストのトップページは，ほぼ完全にテンプレート化しており，トップページから取得できるブロックの差分はほとんどなかった．このため，ブログのトップページが高類似度となってしまい，HTML 構造だけではスプログ検出ができなかった．

3 大規模スプログ収集

前節の結果をふまえて，同一作成者によって大量生成されたスプログのサンプル数が十分でなかった F 社に対して，HTML 構造の類似性を用いて，大規模にスプログ収集を行った．手法としては，500 ブログサイトに対して，52 万ブログサイトのうちまだスプログ・非スプログのラベル付けを行っていないブログサイトとの $Rdiff$ を求め，$Rdiff$ が 0.3 以下のものに対してラベル付けを行った[2]．ここで，F 社の $Rdiff$ の最小値の分布を図 2 に示す．ただし，スプログとの $Rdiff$ の最小値の分布を図 2 (a) に，非スプログとの $Rdiff$ の最小値の分布を図 2 (b) に，それぞれ示す．図 2 (a) より，スプログとの $Rdiff$ の値が小さい範囲には，非スプログは存在せず，大量生成型スプログと単発スプログしか存在しない．また，図 2 (b) より，非スプログとの $Rdiff$ の最小値が 0.3 以下の範囲にはスプログが少ないことが分かる．500 ブログサイトでは，$Rdiff$ が 0.2 以下のブログサイトが 3 サイトしかなかった．しかし，図 2 (a) に示すように，本手法により，既知スプログに類似したスプログを高い密度で収集することができた．

[2]前節では，類似スプログが十分収集されたか否かの基準として，$AvMinDF_{10}(s,T)$ を用いたが，本節では，類似スプログの可能性のあるブログサイトをできるだけ多く収集することが主目的のため，類似スプログが十分収集されたか否かの判断基準を緩めて，$AvMinDF_{10}(s,T)$ ではなく $Rdiff$ の最小値を用いた．

4 おわりに

本論文では，ブログホスト会社 10 社のうち 6 社について，HTML 構造の類似性を用いた教師なし学習により，高適合率でスプログ検出が可能であることを実証した．同一作成者によって大量生成されたスプログのサンプル数が十分でなかった F 社に対しては，サンプルの規模を大きくすることにより，高適合率でスプログ検出が可能であった 6 社と同様の結果が得られることを確認した．非スプログ同士の HTML 構造が類似するため，上記の方式の適用が難しい 2 社 (L 社および Yh 社) については，今後，共通テンプレート部分および非テンプレート部分を分離した上で，HTML 構造の類似性を測定する方式を検討する予定である．

参考文献

[1] 片山太一, 芳中隆幸, 宇津呂武仁, 河田容英, 福原知宏. スプログ検出における HTML 構造の類似性の有効性の評価. 情報処理学会研究報告, Vol. 2009, No. (2009–DBS–149), 2009.

[2] 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. データ工学と情報マネジメントに関するフォーラム—DEIM フォーラム— 論文集, 2010.

[3] P. Kolari, T. Finin, and A. Joshi. Spam in blogs and social media. In *Tutorial at ICWSM*, 2007.

[4] Y. Sato, T. Utsuro, T. Fukuhara, Y. Kawada, Y. Murakami, H. Nakagawa, and N. Kando. Analyzing features of Japanese splogs and characteristics of keywords. In *Proc. 4th AIRWeb*, pp. 33–40, 2008.