DEIM Forum 2010 B4-2

HTML構造を利用した類似スパムブログの収集

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

あらまし　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パム (スパムブログ，スプログ) のうち，特に，同一のスパムブ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を自動収集する手法を提案する．提案手法においては，ブログの HTML ファイルにおけるブロック構造の類似性を用いる．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，同一のスパムブ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．この傾向を主要ブログホスト 10 社で検証し，半数以上のホストに対してスプログを自動収集する手法について提案する．

キーワード　スパムブログ，HTML 構造，DOM，スパムブログ検出

Automatic Collection of Splogs with High Similarities based on HTML Structures

Taichi KATAYAMA†, Takayuki YOSHINAKA††, Takehito UTSURO†, Yasuhide KAWADA†††, and Tomohiro FUKUHARA††††

† Graduate School of Systems and Information Engineering, University of Tsukuba
†† Graduate School of Engineering, Tokyo Denki University
††† Navix Co., Ltd.
†††† Research into Artifacts, Center for Engineering, University of Tokyo

Abstract　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 PageRank of target sites. Among those splogs, this paper focuses on detecting a group of splogs which are estimated to be created by an identical spammer. We especially show that similarities of html structures among those splogs created by an identical spammer contribute to improving the performance of splog detection. In measuring similarities of html structures, we extract a list of DOM (Document Object Model) elements (minimum unit of content) from the DOM tree of an html document. We show that the html documents of splogs estimated to be created by an identical spammer tend to have similar DOM trees and this tendency is quite effective in splog detection.

Key words　spam blog, HTML Structures, DOM, spam blog detection

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* [1]，*kizasi.jp*，

blogWatcher [2] 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．*Blogwise* もまた多言語ブログ記事の分析を行っている．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 [3]〜[7]．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．[5] 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 [8]．このことから, [4], [9] 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 [5]〜[7]．いくつかの既存研究 [5]〜[7] はスプログの重要な特性を報告している．[6] 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, [5], [7] 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, [5], [9]〜[12] では，スプログを機械的に特定し，排除する技術について報告している．

上記の既存研究とは異なり，本論文では，スプログ検出においてスプログの HTML 構造の類似性を利用する [13], [14]．この手法では，HTML 文書の DOM ツリーから DOM 系列を抽出し，DOM 系列の類似性を測定することにより，同一作成者によって作成された多数のスプログを検出する．実際に，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．この傾向を主要ブログホスト 10 社で検証した結果，10 社のうち 6 社は高適合率でスプログ検出が可能であったが，残りの 4 社は，高適合率でのスプログ検出は困難であった．また，高適合率でスプログ検出が困難であった 4 社の傾向は，2 種類に大別することができる．4 社のうちの 2 社においては，非スプログの HTML 構造が類似するため，HTML 構造の類似性を用いたスプログ検出が困難であった．一方，残りの 2 社においては，既知のスプログに対して，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を大規模に自動収集する実験を行うことにより，高適合率でスプログ検出が可能となった．

2. スプログ・非スプログデータセット

日本語ブログ収集にあたり，中国語，日本語，韓国語，英語のブログ記事を収集している KANSHIN システム [15] を利用する．このシステムは各言語のブログサイトのリストを持っている．このリストより，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．2004 年 3 月から 2009 年 5 月までに KANSHIN システムに蓄積された日本語ブログサイト数は，約

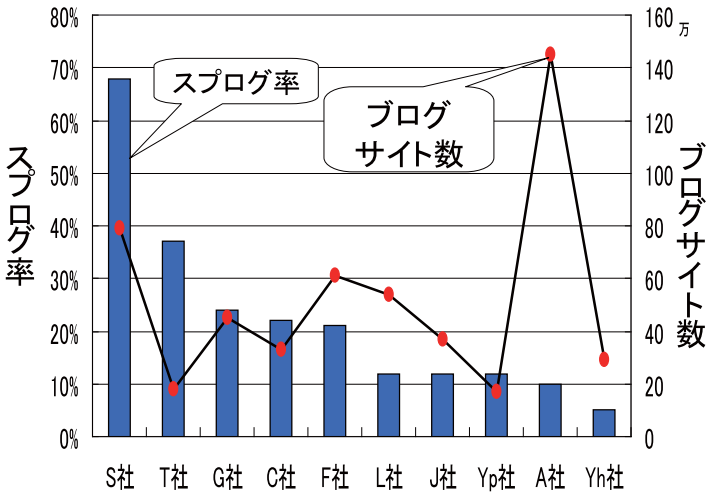


図 1　ホストごとのブログサイト数およびスプログ率

620 万ブログサイトである．そのうち，主要ブログホスト 10 社に限定すると約 520 万ブログサイトである．本論文では，主要ブログホスト 10 社に限定して行う．まず，約 520 万ブログホストのうち，10 分の 1 の 52 万ブログサイトのトップページを取得した．次に，各ホストごとに 500 ブログサイトを無作為に選択し，スプログ・非スプログのラベル付けを行った．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記載する．スプログ率は以下の式で定義した．

$$\text{スプログ率 (ホストごと)} = \frac{\text{スプログ数}}{\text{ホストごとのブログサイト数}}$$

ラベル付けに対しては, [16] で提案された基準によって，スプログ/非スプログを判定する．また，日本語スプログ/非スプログデータセットの中で，テキストコンテンツやブラウザで見た際のフレーム構造が類似しているスプログを，同一作成者が自動生成している「大量生成型」のスプログ [16] として同定し，大量生成型スパマー ID を付与している．それ以外のスプログを「単発」スプログとする．

3. ブログの HTML 構造の類似性の測定

3.1 HTML ファイルからの DOM 系列の抽出

本論文では, [17] で提案されたブロック抽出の方式をふまえて，HTML 文書から DOM 系列を抽出する(注1)(注2)．

まず，図 2 に示すように，HTML 文書 s 中の全ての HTML タグを木構造で表現する．次に，この HTML タグの木構造に

(注1)：[17] も含めて [18]〜[20] 等，HTML 文書からコンテンツを抽出する研究の多くは，自動で主要なコンテンツを抽出する手法に焦点をあてている．その中で, [17] の手法は，HTML 構造の差異を測る際に教師データを必要とせず，しかも主要でない補足的なコンテンツ間の差異の測定への適用も比較的容易であることから，本論文の目的に最も適していた．

(注2)：スプログ検出において HTML 構造の類似度を使用するという基本的な考え方は, [21] においても用いられている．しかし, [21] は，HTML 構造の類似度の計算手法のみにとどまっており，類似度を用いたウェブスパムの検出の評価は行っていない．また，我々の手法と比較すると，HTML タグを用いた類似度尺度の粒度が相対的に粗いと言える．

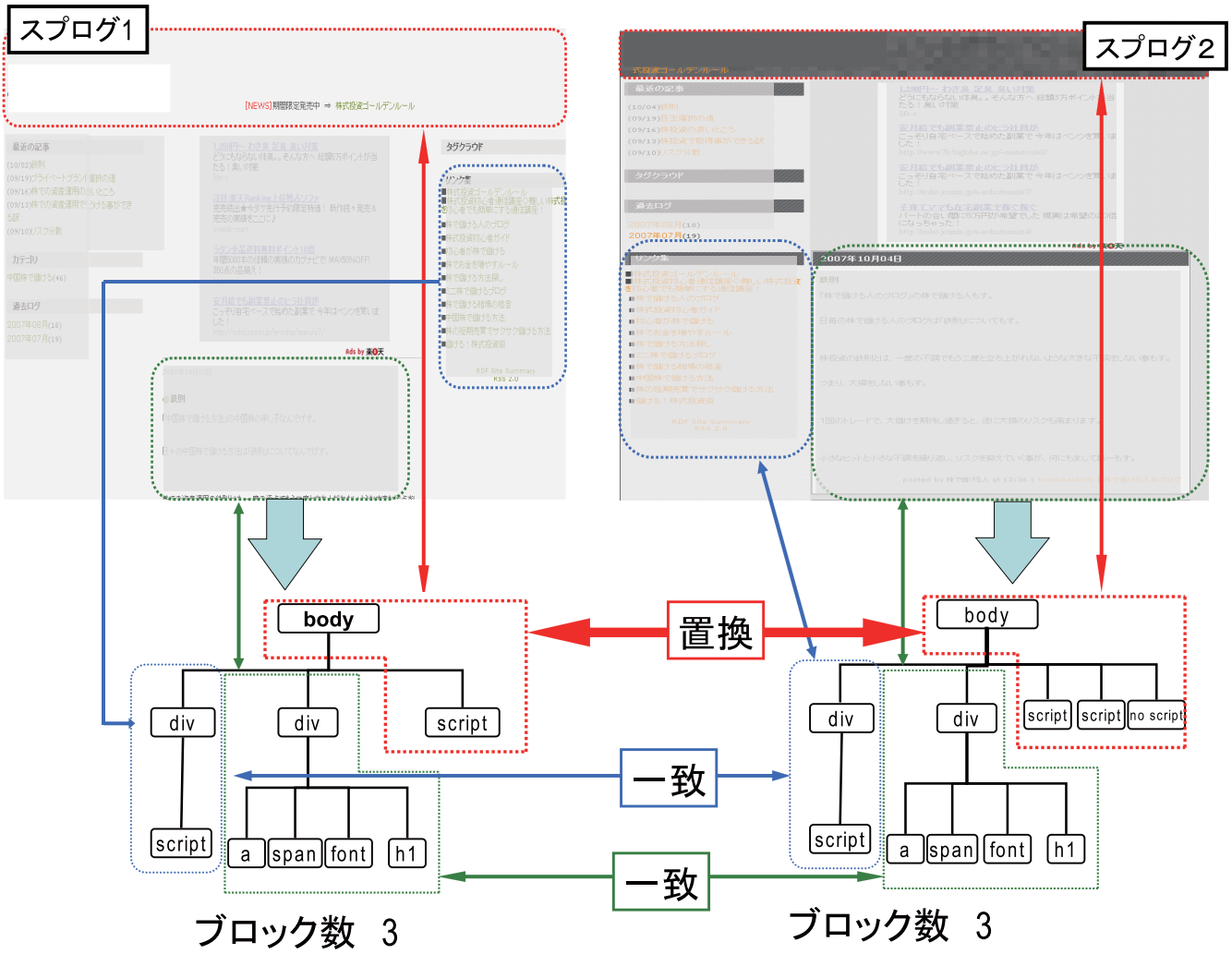


図 2　HTML 文書からの DOM 系列抽出および DOM 系列差分算出の例

対して，ブロックレベル要素として用いられるタグのうち，P タグおよび DIV タグによって木構造を分割し，これらのタグの下位にあるタグを取り込むことによって，個々のブロックを構成する．ここで，一般に，ブロックレベル要素としては，P タグおよび DIV タグ以外のタグも用いられるが，本論文では，簡単化のために，P タグおよび DIV タグに限定する．また，[17] と同様に，BODY タグも，P タグおよび DIV タグと同様に扱い，BODY タグの位置において，HTML タグの木構造の分割を行う．さらに，[17] では，ブラウザにレンダリングされない SCRIPT と STYLE の二タグ及びその下位ノードはブロック内に含めないとしているが，本論文では，ブロックの中身の詳細を区別するために，これらのタグ以下もブロック内に含める．次に，ブロックにまとめあげられた HTML タグの木構造を横型探索することにより，ブロックのリスト構造を形成し，HTML 文書 s の DOM 系列 $dm(s)$ とする．

3.2　DOM 系列の差分の割合

HTML 文書 s および t に対して，それぞれから抽出された DOM 系列 $dm(s)$，および $dm(t)$ の差分を DP マッチングによって求める．DP マッチングの際，挿入および削除のコストを 1，置換のコストを 2 として，DP マッチングにより求まる編集距離 (レーベンシュタイン距離) を $edit\ distance\ (dm(s), dm(t))$ とする．次に，抽出された DOM 系列 $dm(s)$ の要素数を $|dm(s)|$ とし，以下の式で s，t の DOM 系列の差分の割合 $Rdiff(s,t)$ を計算する．

$$Rdiff(s,t) = \frac{edit\ distance\ (dm(s), dm(t))}{|dm(s)| + |dm(t)|}$$

次に，スプログもしくは非スプログの HTML 文書の集合 S および T の間で，HTML 文書 $s \in S$ および $t \in T$ の間の DOM 系列の差分の割合を求め，その分布を分析する．そのためにまず，HTML 文書 $s \in S$ に対して，HTML 文書集合 T の要素 $t \in T$ との間で，差分の割合 $Rdiff(s,t)$ が最も小さい k 個を

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 0.2〜0.3	$AvMinDF_{10}$ の上限値が 0.05〜0.15	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 構造が似てしまう
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	適合率	再現率
C 社	94%	32%
S 社	100%	23%
T 社	95%	74%
A 社	100%	22%
G 社	100%	65%
Yp 社	83%	17%

求め，その差分の割合の平均値を $AvMinDF_k(s,T)$ とする．

$AvMinDF_k(s,T)$ ＝ T 中で，$Rdiff(s,t \in T)$ の値が最も小さい k 個の t に対する $Rdiff(s,t)$ の平均値

4. HTML 構造の類似性を用いたスパムブログ収集

4.1 500 ブログサイトの DOM 系列の差分の割合の分布

10 社のブログホスト会社ごとの 500 ブログサイトに対して，前節で導入した $AvMinDF_k(s,T)$ (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 に書かれているように 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．$AvMinDF_{10}$ の上限値が 0.15〜0.3 であった C 社，S 社，T 社の $AvMinDF_{10}$ の分布を図 3 に，$AvMinDF_{10}$ の上限値が 0.05〜0.15 であった A 社，G 社，Yp 社の $AvMinDF_{10}$ の分布を図 4 に示す．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．C 社，S 社，T 社の 3 社と比べて，A 社，G 社，Yp 社の 3 社では，ブログパーツの配置が限定されているため，ユーザーの手が加えられるところが少なく，DOM 系列での違いが出にくいため，全体的に $AvMinDF_{10}(s,T)$ の値が小さくなってしまった．

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．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 F 社，J 社の $AvMinDF_{10}$ の分布を図 5 に，非スプログ同士の HTML 構造が似てしまう L 社，Yh 社の $AvMinDF_{10}$ の分布を図 6 に示す．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，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ている．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．

4.2 52 万ブログサイトに対する類似スパムブログの収集

前節の結果より，高適合率でスプログ検出が可能であった C 社，S 社，および，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，大規模にスパムブログ収集を行った．手法としては，500 ブログサイトでの $AvMinDF_{10}(s,T)$ の値が 0.3 以上の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(注3)(注4)．

ここで，F 社を対象とした場合の，$Rdiff$ の最小値の分布を図 7 に示す．ただし，スプログとの $Rdiff$ の最小値の分布を図 7 (a) に，非スプログとの $Rdiff$ の最小値の分布を図 7 (b) に，それぞれ示す．図 7(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 7(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．C 社を対象とした場合の，$Rdiff$ の最小値の分布を図 8 に，S 社を対象とした場合の，$Rdiff$ の最小値の分布を図 9 に示す．C 社，S 社の場合における $Rdiff$ の最小

(注3)：500 ブログサイトでの $AvMinDF_{10}(s,T)$ の値が 0.3 以下のものについては，すでに，同一作成者が作成した類似スプログが 10 サイト以上収集済であると考え，500 ブログサイトでの $AvMinDF_{10}(s,T)$ の値が 0.3 以下のものは，52 万ブログサイトを対象とした類似スプログ収集の対象からは除外した．

(注4)：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．

表 3　52 万ブログサイトを対象とする類似スパムブログ収集の性能

ブログホスト会社	スプログとの $Rdiff$ の最小値 ≦0.15 のブログサイト		非スプログとの $Rdiff$ の最小値 ≦0.3 のブログサイト	
	ブログサイト数	スプログ率	ブログサイト数	非スプログ率
C 社	91	96%	1625	89%
S 社	143	99%	505	50%
F 社	67	99%	427	88%

値の分布も，ほぼ同様の傾向であった．そこで，スプログとの $Rdiff$ の最小値の上限値を 0.15 とし，非スプログとの $Rdiff$ の最小値の上限値を 0.3 として，52 万ブログサイトより収集されたブログサイトの数，スプログ率 (スプログとの $Rdiff$ の最小値 ≦ 0.15 の場合)，および，非スプログ率 (非スプログとの $Rdiff$ の最小値 ≦ 0.3 の場合) を表 3 に示す．この結果のうち，特にスプログ率については，非常に高い値を達成しており，52 万ブログサイトから，高密度で類似スプログを自動収集できていることが分かる．

5. 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．また，類似スプログ収集においては，既知スプログに類似したスプログを高い密度で収集可能であることを示した．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．[14] では，HTML 構造の類似性素性を素性の一つとして教師あり学習を行ったところ，主要 10 社のうち 2 社では HTML 構造の類似性素性を追加したことで精度が向上したことが確認されているため，本論文で高適合率でスプログ検出が可能であった 6 社全てで HTML 構造の類似性素性の有効性を評価する必要がある．また，我々は，[22] において，能動学習の枠組みを用いてスプログを自動収集する手法を提案しているが，この手法においても，本論文で提案した HTML 構造の類似性素性が効果的である可能性が高いので，その有効性を検証する必要性がある．

文　献

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

[2] T. Nanno, T. Fujiki, Y. Suzuki, and M. Okumura. Automatically collecting, monitoring, and mining Japanese weblogs. In *WWW Alt. '04: Proc. 13th WWW Conf. Alternate Track Papers & Posters*, pp. 320–321, 2004.

[3] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proc. 1st AIRWeb*, pp. 39–47, 2005.

[4] *Wikipedia, Spam blog.* `http://en.wikipedia.org/wiki/Spam_blog`.

[5] P. Kolari, A. Joshi, and T. Finin. Characterizing the splogosphere. In *Proc. 3rd Ann. Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2006.

[6] C. Macdonald and I. Ounis. The TREC Blogs06 collection : Creating and analysing a blog test collection. Technical Report TR-2006-224, University of Glasgow, Department of Computing Science, 2006.

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

[8] *Wikipedia, Ping (blogging)*. `http://en.wikipedia.org/wiki/Ping_(blogging)`.

[9] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. L. Tseng. Splog detection using self-similarity analysis on blog temporal dynamics. In *Proc. 3rd AIRWeb*, pp. 1–8, 2007.

[10] 石田和成. 共起クラスターシードと連鎖的抽出にもとづくスパムブログのフィルタリング. Web とデータベースに関するフォーラム (WebDB Forum)2008 論文集. 情報処理学会, 2008.

[11] P. Kolari, T. Finin, and A. Joshi. SVMs for the Blogosphere: Blog identification and Splog detection. In *Proc. 2006 AAAI Spring Symp. Computational Approaches to Analyzing Weblogs*, pp. 92–99, 2006.

[12] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *Proc. 1st AIRWeb*, 2005.

[13] 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.

[14] T. Katayama, T. Yoshinaka, T. Utsuro, Y. Kawada, and T. Fukuhara. Detecting splogs using similarities of splog HTML structures. In *Proc. 4th Inter. Conf. on Ubiquitous Information Management and Communication*, pp. 256–263, 2010.

[15] T. Fukuhara, A. Kimura, Y. Arai, T. Yoshinaka, H. Masuda, T. Utsuro, and H. Nakagawa. KANSHIN: A cross-lingual concern analysis system using multilingual blog articles. In *Proc. 1st INGS2008*, pp. 83–90, 2008.

[16] 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.

[17] 吉田光男, 山本幹雄. 教師情報を必要としないニュースページ群からのコンテンツ自動抽出. 日本データベース学会論文誌, Vol. 8, No. 1, pp. 29–34, 2009.

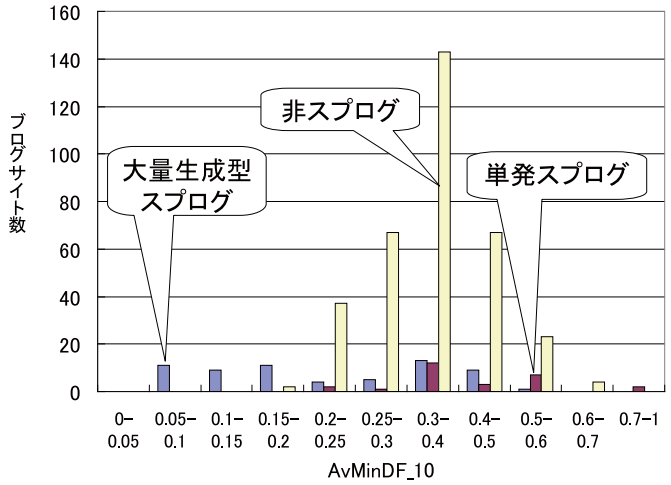
[18] S.-H. Lin and J.-M. Ho. Discovering informative content blocks from Web documents. In *Proc. 8th SIGKDD*, pp. 588–593, 2002.

[19] S. Debnath, P. Mitra, N. Pal, and C. L. Giles. Automatic identification of informative sections of Web pages. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 9, pp. 1233–1246, 2005.

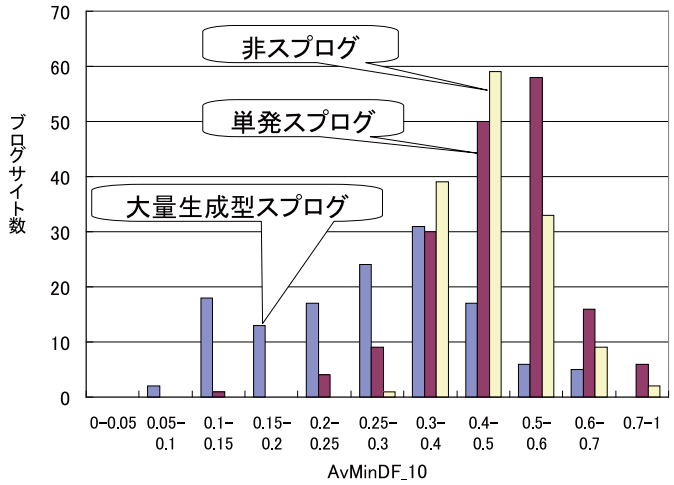
[20] L. Bing, Y. Wang, Y. Zhang, and H. Wang. Primary content extraction with mountain model. In *Proc. 8th IEEE CIT*, pp. 479–484, 2008.

[21] T. Urvoy, T. Lavergne, and P. Filoche. Tracking Web spam with hidden style similarity. In *Proc. 2nd AIRWeb*, pp. 25–30, 2006.

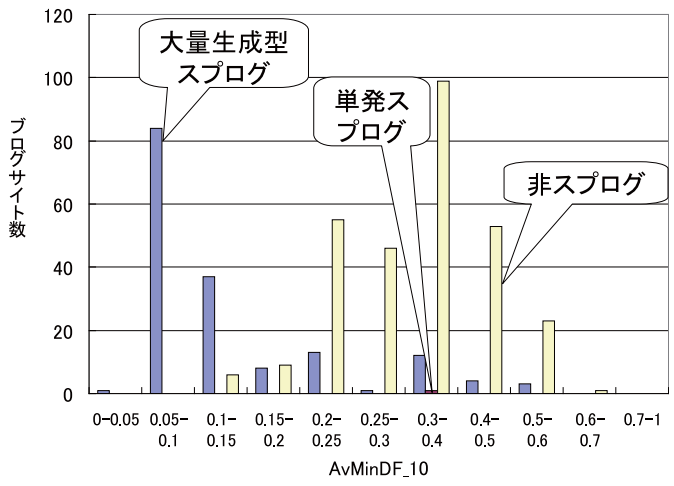
[22] T. Katayama, Y. Sato, T. Utsuro, T. Yoshinaka, Y. Kawada, and T. Fukuhara. An empirical study on selective sampling in active learning for splog detection. In *Proc. 5th AIRWeb*, pp. 29–36, 2009.



(a) C 社

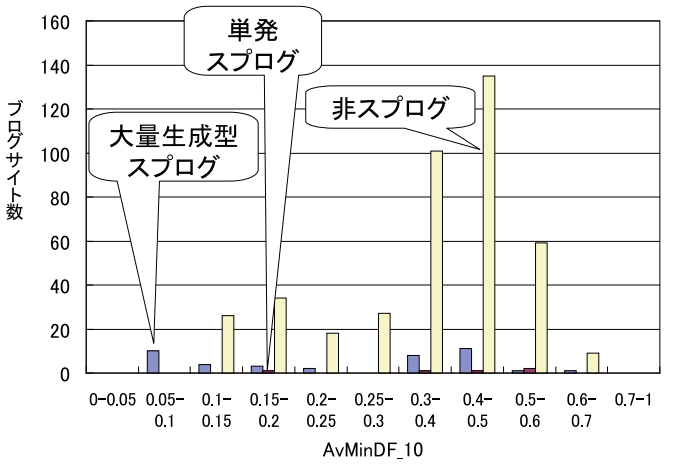


(b) S 社

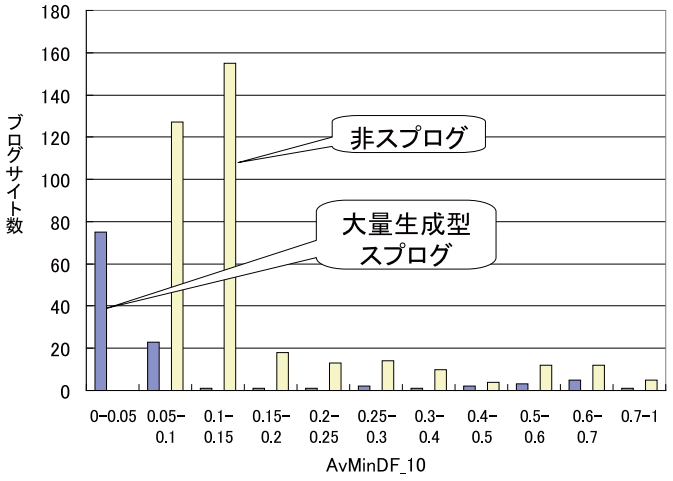


(c) T 社

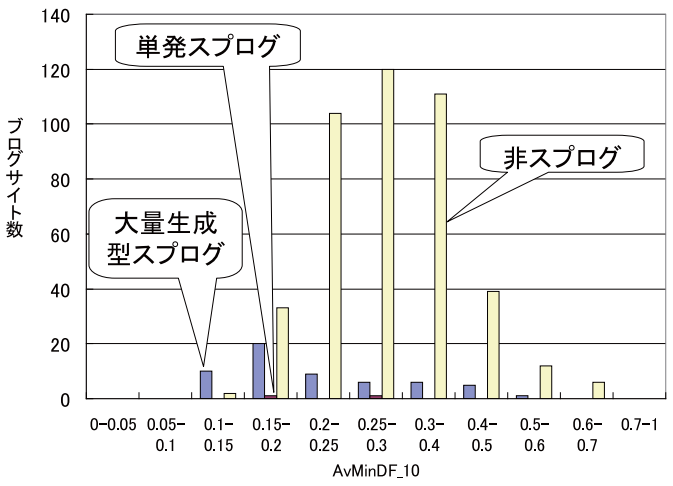
図 3　500 ブログサイトの DOM 系列の差分の割合の分布 (HTML 構造のみを用いたスプログ検出が高適合率で可能，$AvMinDF_{10}$ の上限値が 0.15〜0.3 の 3 社)



(a) A 社

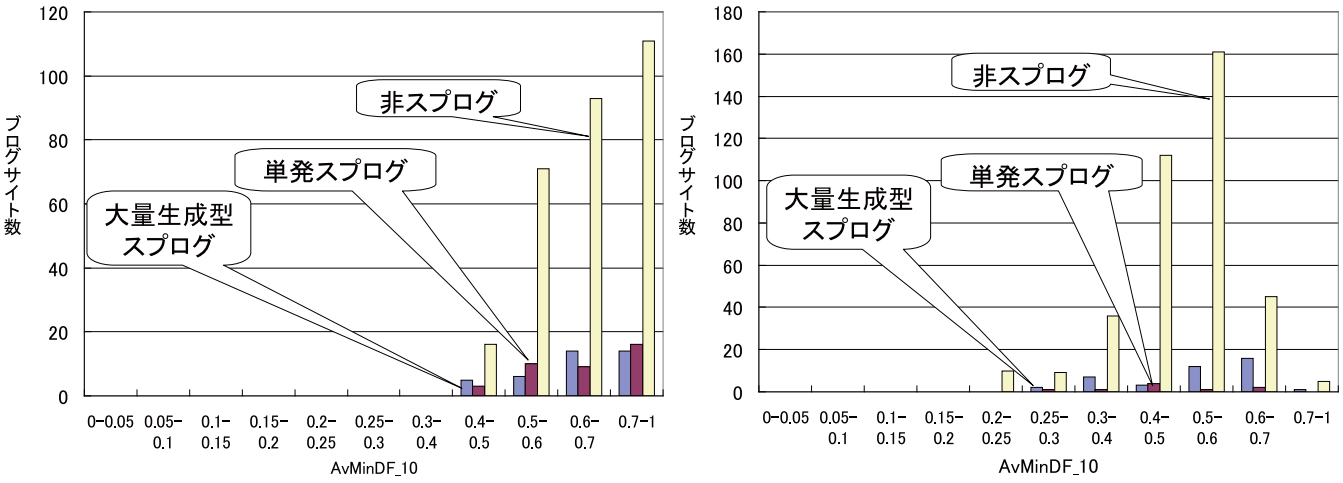


(b) G 社



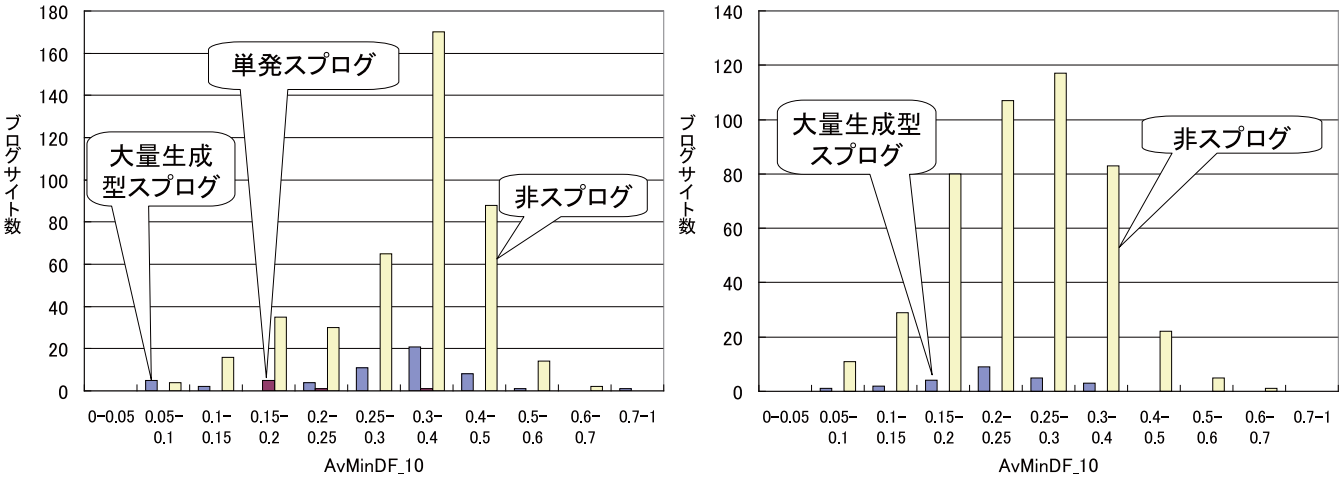
(c) Yp 社

図 4　500 ブログサイトの DOM 系列の差分の割合の分布 (HTML 構造のみを用いたスプログ検出が高適合率で可能，$AvMinDF_{10}$ の上限値が 0.05〜0.15 の 3 社)



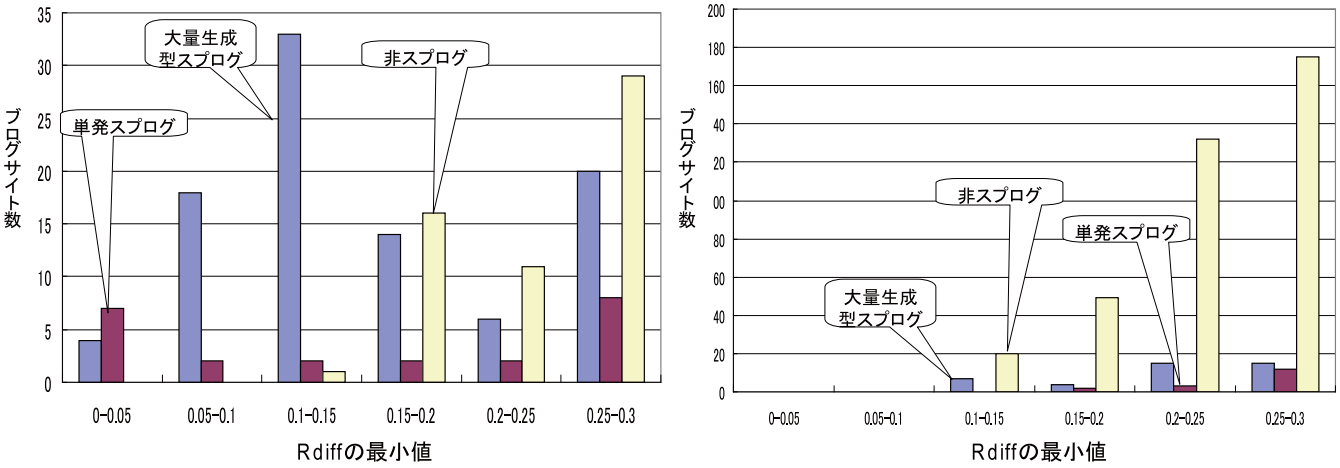
(a) F 社　　(b) J 社

図 5　500 ブログサイトの DOM 系列の差分の割合の分布 (HTML 構造のみを用いたスプログ検出は高適合率では困難，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 2 社)



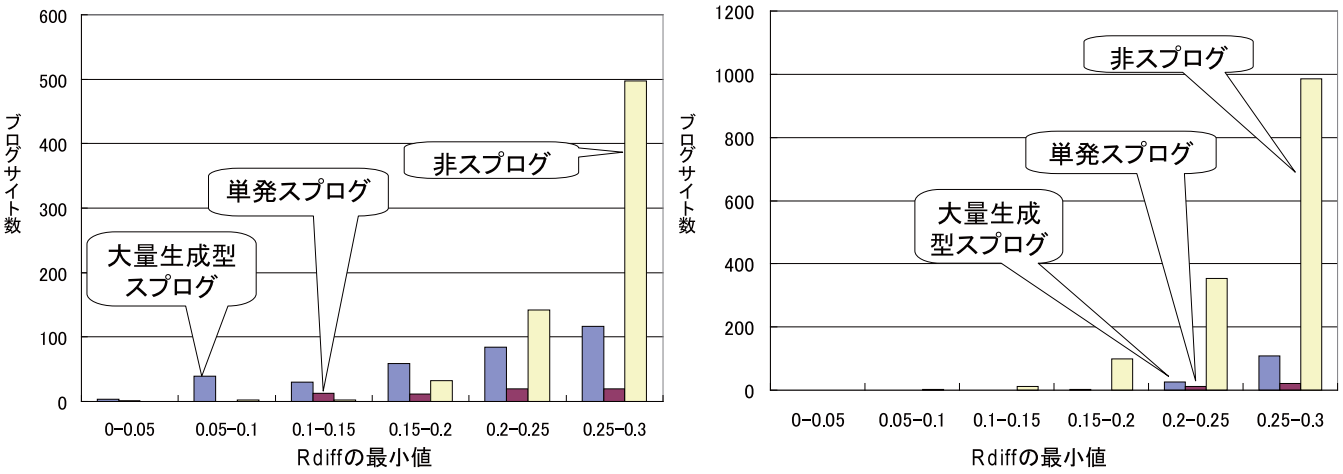
(a) L 社　　(b) Yh 社

図 6　500 ブログサイトの DOM 系列の差分の割合の分布 (HTML 構造のみを用いたスプログ検出は高適合率では困難，非スプログ同士が似てしまう 2 社)



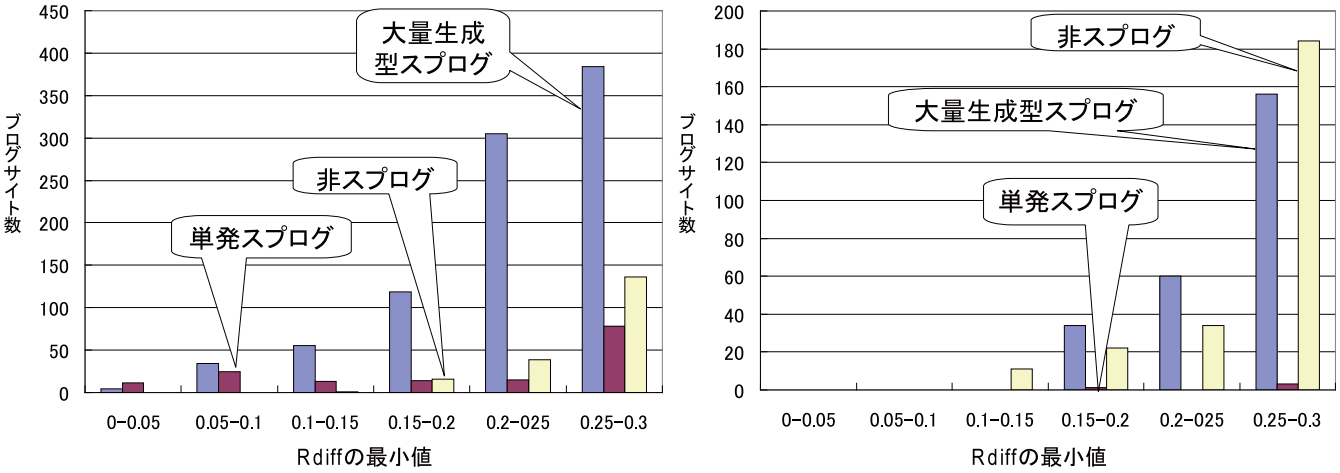
(a) スプログとの $Rdiff$ の最小値　　(b) 非スプログとの $Rdiff$ の最小値

図 7　52 万ブログサイトを対象とする $Rdiff$ の最小値の分布 (F 社)



(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 8　52 万ブログサイトを対象とする $Rdiff$ の最小値の分布 (C 社)



(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 9　52 万ブログサイトを対象とする $Rdiff$ の最小値の分布 (S 社)