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

片山太一[†1] 森尻惇宜史[†2] 石井聡一[†3]
宇津呂武仁[†1] 河田容英[†4] 福原知宏[†5]

本論文では，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) について，スプログの HTML 構造の類似性およびアフィリエイト ID という異なる二種類の手がかりの特性を分析する．まず，既知のスプログに対して HTML 構造が類似するブログサイトを大規模に収集することにより，既知スプログに類似するスプログが高密度で自動収集できることを示す．また，高類似度なスプログに対して，作成者 (スパマー) の同一性の判定を人手で行い，同一のスパマーが作成したと推定されるスプログについても，高密度で収集することができることを示す．一方，単一のアフィリエイト ID が多数のブログサイトによって共有されている場合には，そのアフィリエイト ID はスパムである可能性が高い．しかし，現時点において，主要な ASP (Affiliate Service Provider) にわたって，網羅的にアフィリエイト ID を抽出するのは容易ではなく，同一のスパマーが作成したと推定されるスプログにおいても，アフィリエイト ID が抽出できるブログサイトの割合が限定的であることを示す．さらに，アフィリエイト ID を共有する複数のスプログの間で，HTML 構造が類似するスプログが一定の割合で存在し，アフィリエイト ID と HTML 構造の類似性の間に相関があることを示す．

Analyzing Splogs based on Similarities of HTML Structures and Affiliate

Taichi Katayama,[†1] Akihito Morijiri,[†2] Soichi Ishii,[†3]
Takehito Utsuro,[†1] Yasuhide Kawada[†4]
and Tomohiro Fukuhara[†5]

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. Among those splogs, this paper focuses on detecting a group of splogs which are estimated to be created by an identical spammer. 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, as well as in identifying spammers. We next show that it is not easy to achieve high coverage in extracting affiliate IDs, even though blog sites sharing an identical affiliate ID tend to be splogs. Finally, we show that certain percentages of splogs which share an identical affiliate ID have similar HTML structures, and thus similarities of splog HTML structures and affiliate have certain correlation.

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*，*kizasi.jp*，*blogWatcher* 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 [2),8),9),11)]．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．文献 9) 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，文献 10) に述べられてい

†1 筑波大学大学院システム情報工学研究科
Graduate School of Systems and Information Engineering,University of Tsukuba
†2 筑波大学理工学群工学システム学類
College of Engineering Systems, School of Science and Engineering, University of Tsukuba
†3 東京電機大学大学院未来科学研究科
Graduate School of Science and Technology for Future Life, Tokyo Denki University
†4 (株) ナビックス
Navix Co., Ltd.
†5 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology

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 8)，9)，11) はスプログの重要な特性を報告している．文献 11) では，TRECBlog06 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，文献 8)，9) 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，文献 4)，7)，10)，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，スプログの HTML 構造の類似性およびアフィリエイト ID という異なる二種類の手がかりを用いてスプログの特性を分析する．スプログにおいては，一人の作成者が大量のスプログを機械的に大量生成していると考えられる[6),13)]．また，我々のこれまでの分析[6)] においては，これらのスプログにおいて，HTML 構造が類似する傾向があることがわかっている．ここで，本論文では，既知のスプログに対して HTML 構造が類似するブログサイトを大規模に収集することにより，既知スプログに類似するスプログが高密度で自動収集できることを示す (3.3.1 節)．また，高類似度なスプログに対して，作成者 (スパマー) の同一性の判定を人手で行い，同一のスパマーが作成したと推定されるスプログについても，高密度で収集することができることを示す (3.3.2 節)．一方，一人のスプログ作成者が機械的に生成した複数スプログにおいて，その作成者が所有する単一のアフィリエイト ID を使用しているという現象が観測されている[5)]．実際に，単一のアフィリエイト ID が多数のブログサイトによって共有されている場合には，そのアフィリエイト ID はスパムである可能性が高い[5)]．しかし，現時点において，主要な ASP (Affiliate Service Provider) にわたって，網羅的にアフィリエイト ID を抽出するのは容易ではなく，同一のスパマーが作成したと推定されるスプログにおいても，アフィリエイト ID が抽出できるブログサイトの割合が限定的であることを示す (3.3.3 節)．さらに，アフィリエイト ID を共有する複数のスプログの間で，HTML 構造が類似するスプログが一定の割合で存在し，アフィリエイト ID と HTML 構造の類似性の間に相関があることを示す (4.2 節)．

2. ブログの HTML 構造の類似性の測定

2.1 HTML ファイルからの DOM 系列の抽出

本論文では，文献 14) 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．まず，図 1 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，文献 14) 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，文献 14) 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．

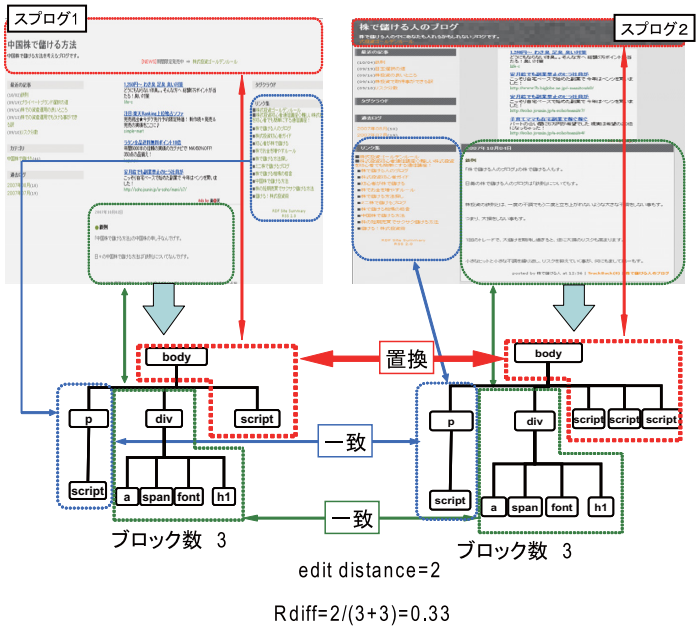


図 1 HTML 文書からの DOM 系列抽出および DOM 系列差分算出の例

2.2 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

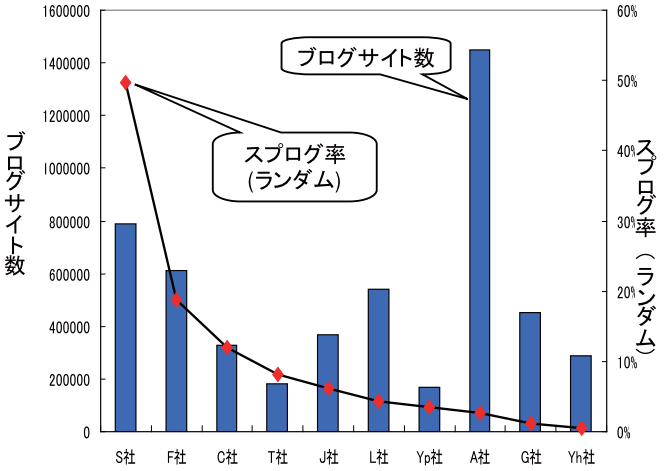


図 2 ホストごとのブログサイト数および「スプログ率 (ランダム)」

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 $\mathrm{Rdiff}(s,t)$ を計算する．

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

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，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 $\mathrm{Rdiff}(s,t)$ が最も小さいものを求め，その差分の割合の最小値を $\mathrm{MinDF}(s,T)$ とする．

$$\mathrm{MinDF}(s,T) = \min_{t \neq s} \mathrm{Rdiff}(s, t \in T)$$

3. HTML 構造の類似性を用いて収集したスプログの分析

3.1 初期スプログ集合

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[1] 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．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 620 万ブログサイトである．これを，主要ブログホスト 10 社に限定すると約 520 万ブログサイトとなる．まず，各ホストについて，350~600 のブログサイトを無作為に選択し，文献 13) の基準に基づいてスプログ・非スプログの判定を行った．そして，以下の「スプログ率 (ランダム)」を算出した．

表 1 初期スプログ集合 $SP_{seed}(H)$ のサイズ

ブログホスト H	$\lvert SP_{seed}(H) \rvert$	17 万ブログサイト中に $\mathrm{Rdiff} \leq k$ となるブログサイト b が存在するスプログ $s(\in SP_{seed}(H))$ の数					
		$k = 0.05$	$k = 0.1$	$k = 0.15$	$k = 0.2$	$k = 0.25$	$k = 0.3$
S 社	289	13	34	50	78	106	136
F 社	100	11	16	21	28	35	45
C 社	100	17	31	46	61	68	83

$$\text{スプログ率 (ランダム)} = \frac{\text{スプログ数}}{\text{スプログ・非スプログ判定対象のブログサイト数}}$$

図 2 に，主要 10 社のブログホストのブログサイト数，および，「スプログ率 (ランダム)」を示す．

以下，本論文では，図 2 の「スプログ率 (ランダム)」において，上位 3 社のブログホスト会社に限定して分析を行う．まず，各ホストのスプログ数が 100 以上になるように，無作為にブログサイトを選択してスプログ・非スプログの判定を行い，スプログを追加した．以上の手順で収集したスプログ集合を，各ブログホスト H についての初期スプログ集合 $SP_{seed}(H)$ とする．各初期スプログ集合のサイズを表 1 左側に示す．

3.2 類似スプログの収集手順

次に，前節の初期スプログ集合 $SP_{seed}(H)$ を既知スプログ集合として，2 節で述べた HTML 構造の類似性を用いて類似ブログサイトを収集する手順を以下に述べる．本手法では，ブログホストごとに HTML タグの使い方の傾向やテンプレートが異なるため，HTML 構造の類似性の測定をブログホストごとに行う[⋆1]．

分析対象としたブログホスト 3 社のブログサイト数の合計は約 170 万であった．そこで，そのうちの 10%，17 万ブログサイトを無作為に選択し，類似ブログサイト収集の対象とした．

各ホスト H について収集対象としたブログサイトの集合を $B(H)$，そのうちのスプログの集合を $SP(H)$ とする．次に，$B(H)$ の要素 b のうち，既知スプログ集合 $SP_{seed}(H)$ の

⋆1 複数のブログホストでスプログを作成しているスパマーが存在することも十分考えられるが，それらのスパマーは各ブログホストで複数スプログを作成していると考えられるため，本論文では，類似スプログの収集をブログホストごとに行う．

いずれかの要素 s との間で，$k-0.05 \leq \mathrm{Rdiff}(s,b) \leq k$ となるブログサイト b (言い換えれば，$k-0.05 \leq \mathrm{MinDF}(b, SP_{seed}(H)) \leq k$ となるブログサイト b) の集合を $B(H,k)$ と定義する．また，$B(H,k)$ のうちのスプログの集合を $SP(H,k)$ と定義する．

$$B(H,k) = \Big\{ \text{ホスト } H \text{ のブログサイト } b \;\Big|\; b \notin SP_{seed}(H),\ \ k-0.05 \leq \mathrm{MinDF}(b, SP_{seed}(H)) \leq k \Big\}$$

$$SP(H,k) = \Big\{ \text{ スプログサイト } s \;\Big|\; s \in B(H,k) \Big\}$$

逆に，既知スプログ集合 $SP_{seed}(H)$ の要素 s のうち，$B(H)$ のいずれかの要素 b との間で，$k-0.05 \leq \mathrm{Rdiff}(s,b) \leq k$ となるスプログ s の集合を $SP_{seed}(H,k)$ と定義する．

$$SP_{seed}(H,k) = \Big\{ s \in SP_{seed}(H) \;\Big|\; \exists b \in B(H),\ \ b \notin SP_{seed}(H),\ \ k-0.05 \leq \mathrm{Rdiff}(s,b) \leq k \Big\}$$

ここで，$B(H)$ の要素のうち，$\mathrm{MinDF}(b, SP_{seed}(H))$ の値が 0.3 以下となるブログサイトについて，既知スプログ集合の中で HTML 構造が類似するスプログが存在するブログサイトであるとみなして，次節における分析対象とした．なお，既知スプログ集合 $SP_{seed}(H)$ の要素 s のうち，$B(H)$ のいずれかの要素 b との間で，$\mathrm{Rdiff}(s,b) \leq k$ ($k = 0.05,\ 0.1,\ 0.15,\ 0.2,\ 0.25,\ 0.3$) となるスプログ s の数を表 1 右側に示す．これから分かるように，$k = 0.3$ の場合，ブログホスト 3 社のうち 2 社については，既知スプログのうちの半数近くについて，HTML 構造の類似するブログサイトが存在することが分かる．また，残りの 1 社については，その割合は半数を越えていることがわかる．

3.3 収集されたスプログの分析

3.3.1 スプログ率の分析

3.2 節の手順によって，$B(H)$ の要素のうち，$\mathrm{MinDF}(b, SP_{seed}(H))$ の値が 0.3 以下となるブログサイトを収集したところ，その数は，それぞれ，7,787 サイト (S 社)，623 サイト (F 社)，および，7,352 サイト (C 社) となった．このうち，F 社については，全サイトについて，人手によりスプログ・非スプログの判定を行った．一方，残りの 2 社については，各既知スプログ $s(\in SP_{seed}(H))$ ごとに，最大 10 サイトを選択し，人手によりスプログ・非スプログの判定を行った[⋆1] [⋆2]．

次に，$k = 0.05,\ 0.1,\ 0.15,\ 0.2,\ 0.25,\ 0.3$ の各値に対して，収集されたブログサイト集合 $B(H,k)$ を対象として，以下の「スプログ率 (既知スプログと高類似度のブログサイト)」を算出し[⋆3]，その値を図 3 にプロットした．

$$\text{スプログ率 (既知スプログと高類似度のブログサイト)} = \frac{\big|SP(H,k)\big|}{\big|B(H,k)\big|}$$

このスプログ率は，既知スプログ集合 $SP_{seed}(H)$ との間で，

$$k-0.05 \leq \mathrm{MinDF}(b, SP_{seed}(H)) \leq k$$

という高い HTML 構造の類似度を持つブログサイト b を収集した集合におけるスプログ率である．また，比較のために，図 2 に示した「スプログ率 (ランダム)」も示す．図 3 には，$k = 0.05,\ 0.1,\ 0.15,\ 0.2,\ 0.25,\ 0.3$ の各値に対して，収集されたスプログ数 (F 社)，もしくは，その推定値 (S 社および C 社)

$$\text{「スプログ率 (既知スプログと高類似度のブログサイト)」} \times \big|B(H,k)\big|$$

も示す．

図 3 の評価結果に対する分析の要点を以下に示す．

(1) いずれのブログホストにおいても，「スプログ率 (既知スプログと高類似度のブログサイト)」は，「スプログ率 (ランダム)」を大幅に上回っている．このことから，高密度でスプログが収集できていることが分かる．特に，C 社では，MinDF の範囲が 0.05 以下でのスプログ率が極めて高く，F 社では，0.15 以下の範囲，S 社では，0.3 以下全体を通してスプログ率が高い．

(2) 既知スプログ数に対して，新たに収集されるスプログの数が極めて多い．S 社では，MinDF の範囲が 0.15 以下で，50 サイトの既知スプログから，1,160 サイトのスプログが 95%の高い密度で新たに収集できる．MinDF の範囲が 0.3 以下では，136 サイトの既知スプログから，6,824 サイトのスプログが 87%の高い密度で新たに収集できる．一方，F 社では，MinDF の範囲が 0.15 以下で，21 サイトの既知スプログから，118 サイトのスプログが 98%の高い密度で新たに収集できる．また，C 社では，MinDF の範囲が 0.05 以下で，17 サイトの既知スプログから，155 サイトのスプロ

⋆1 最も類似する 5 サイト，および，Rdiff を分散させた 5 サイト，の合計 10 サイト．

⋆2 スプログ数・非スプログ数は，それぞれ，880 サイト・133 サイト (S 社)，346 サイト・282 サイト (F 社)，および，368 サイト・190 サイト (C 社) となった．

⋆3 S 社および C 社については，$B(H,k)$，および，$SP(H,k)$ 中で，スプログ・非スプログの判定を付与したブログサイトにより構成した部分集合 $B'(H,k)$ および $SP'(H,k)$ を用いて，$\frac{|SP'(H,k)|}{|B'(H,k)|}$ により，「スプログ率 (既知スプログと高類似度のブログサイト)」を算出する．

(a) S 社

(b) F 社

(c) C 社

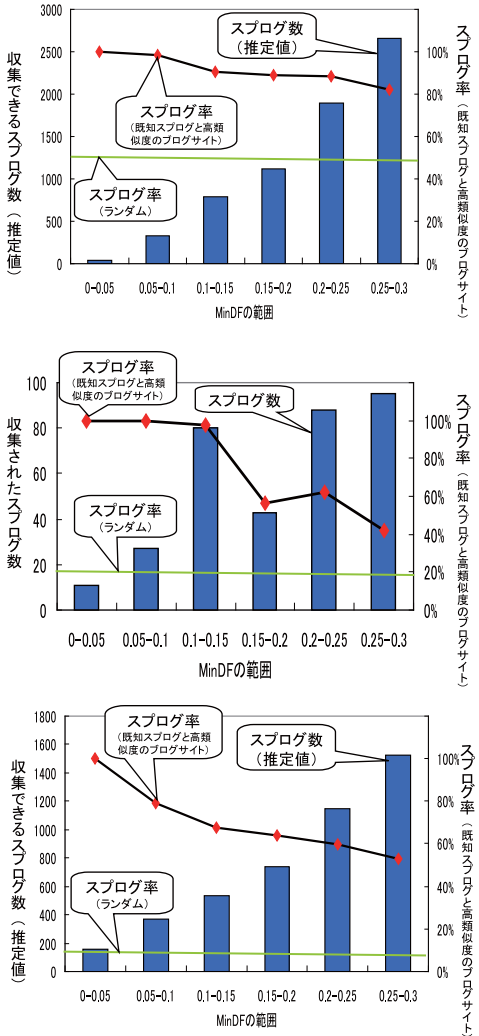


図 3　「スプログ率 (既知スプログと高類似度のブログサイト)」および収集されるスプログ数 (推定値)

グが 100%の高い密度で新たに収集できる[⋆1]．

⋆1 他の 2 社と比較すると，C 社は「スプログ率 (既知スプログとの高類似度のブログサイト)」が低い．C 社においては，ブログホストの制限により，ブログのトップページには，記事の一部のみが掲載されているブログサイトが多く存在した．これらの DOM 系列が似てしまうことにより，Rdiff が小さくなる傾向がみられた．

(a) S 社

(b) F 社

(c) C 社

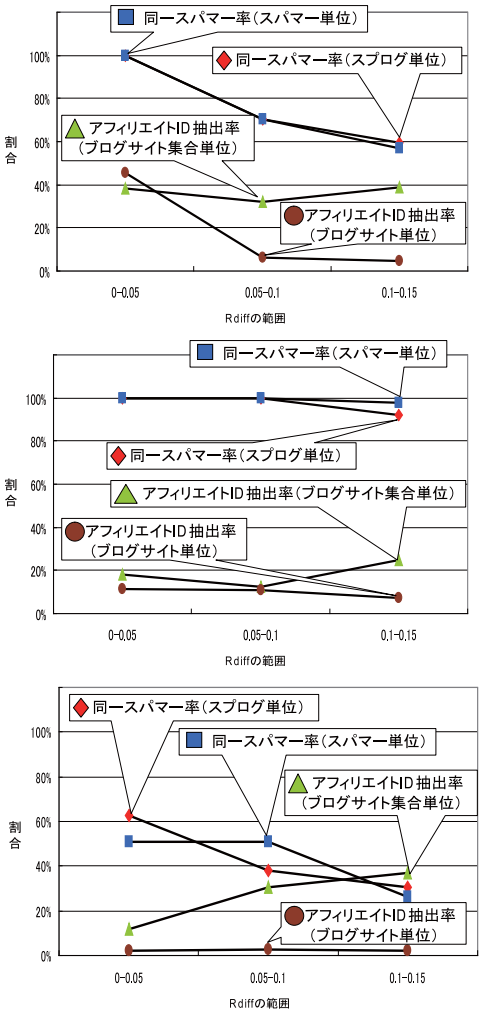


図 4　同一スパマー率およびアフィリエイト ID 抽出率の評価結果

3.3.2　同一スパマーの分析

次に，前節で新たに収集したスプログに対して，以下の基準により，既知スプログの作成

者と同一のスパマーが作成したものかどうかの判定を行った．

(1)　コピー元の文書に対して固有名，同義語の置換を行ったもの．

(2)　ブラウザで表示されるフレーム構造やフレーム内のレイアウトが類似している．

(3)　リンク先の URL が共通である．

ここでは，前節におけるスプログ率の分析結果に基づいて，ブログホスト 3 社とも，一律に，MinDF の範囲が 0.15 以下という条件で新たに収集されたスプログを分析対象とした．

具体的には，$k = 0.05,\ 0.1,\ 0.15$ の各値に対して，既知スプログ集合 $SP_{seed}(H,k)$ の各要素の既知スプログ (合計では，表 1 の $k = 0.15$ の欄より，50 サイト (S 社)，21 サイト (F 社)，および 46 サイト (C 社)) に類似するスプログを収集したもの (189 サイト (S 社)，146 サイト (F 社)，および 159 サイト (C 社)) に対して，同一スパマーによって作成されたか否かを判定した．

この判定結果を集計して評価するための尺度を以下で定義する．まず，$B(H)$ の要素 b のうち，既知スプログ集合 $SP_{seed}(H)$ の要素 s との間で，$k - 0.05 \ \leq\ \mathrm{Rdiff}(s,b)\ \leq\ k$ となるブログサイト b の集合を $B(H,s,k)$ と定義する．

$$B(H,s,k) = \left\{ b \in B(H) \;\middle|\; b \notin SP_{seed}(H),\ \ k - 0.05 \ \leq\ \mathrm{Rdiff}(s,b)\ \leq\ k \right\}$$

また，既知スプログ集合 $SP_{seed}(H)$ の要素 s のうち，新たに収集したブログサイト集合 $B(H,s,k)$ 中に，s の作成者と同一のスパマーにより作成されたと推定できるスプログが複数含まれるものを集めた集合を $SP'_{seed}(H,k)$ と定義する．

$$SP'_{seed}(H,k) \ \ = \ \ \left\{ s \in SP_{seed}(H,k) \;\middle|\; B(H,s,k) \text{ の中に，} s \text{ の作成者と同一の} \right.$$
$$\left. \text{スパマーにより作成されたと推定できるスプログが複数含まれる} \right\}$$

以上をふまえて，新たに収集したスプログ集合に対して，既知スプログのサイト数の単位で算出した同一スパマー率を以下の「同一スパマー率 (スパマー単位)」として定義する．

$$\text{同一スパマー率 (スパマー単位)} = \frac{\left|SP'_{seed}(H,k)\right|}{\left|SP_{seed}(H,k)\right|}$$

同様に，既知スプログ集合 $SP_{seed}(H)$ 中の各既知スプログ s に対して，$B(H,s,k)$ 中に，s の作成者と同一のスパマーにより作成されたと推定できるスプログが複数含まれる場合には，$B(H,s,k)$ の全ての要素が同一スパマーによって作成されたと近似し，新たに収集したスプログ数の単位で算出した同一スパマー率を以下の「同一スパマー率 (スプログ単位)」として定義する．

$$\text{同一スパマー率 (スプログ単位)} = \frac{\displaystyle\sum_{s \in SP'_{seed}(H,k)} \left|B(H,s,k)\right|}{\displaystyle\sum_{s \in SP_{seed}(H,k)} \left|B(H,s,k)\right|}$$

以上の評価値をプロットした結果を図 4 に示す．また，図 4 の評価結果に対する分析の要点を以下に示す[⋆1]．

(1)　F 社については，新たに収集されたスプログのほぼすべてが，既知スプログの作成者と同一のスパマーにより作成されたと判断でき，その特徴が HTML 構造の類似性に現れていた．

(2)　S 社，C 社とも,「同一スパマー率」としては，図 3 におけるスプログ率の評価結果を下回る結果となった．その主要な原因として，既知スプログ集合中に，記事数が数記事程度のスプログが含まれており，これに類似するスプログとして，同様に記事数が数記事程度のスプログが新たに収集されたことが挙げられる．これらのスプログは，通常，作成者が異なっていると判断することが適切であり，本論文の評価においても，同一スパマーによって作成されたものとは判定しなかった．ただし，これらのスプログを除外すれば，同一スパマーによって作成されたと推定されるスプログが一定数収集されている．

3.3.3　アフィリエイト ID の分析

スプログには，通常，アフィリエイトが含まれていると考えられる[3]．また，一人のスパマーが機械的に生成した複数スプログにおいて，その作成者が所有する単一のアフィリエイト ID を使用しているという現象が観測されている[5]．したがって，同一スパマーを特定する有力な手がかりとして，複数のスプログから抽出したアフィリエイト ID の一致の有無が挙げられる．そこで，4 節においては，複数のスプログにおいて共有されているアフィリエイト ID を用いて収集したスプログ集合を対象として，HTML 構造の類似性の分析を行うが，そのための準備として，本節では，3.2 節の手順で収集したスプログ集合において，どの程度のカバレージでアフィリエイト ID の抽出が可能であるかの評価を行う．

まず，4.1 節で詳細に述べるように，本節においては，文献 5) の手順にしたがって，ブロ

⋆1 F 社のあるスプログとの類似度が 0 のものが 5 サイト，0.057 のものが 3 サイト収集できた．これらを人手で分析した結果，全て同一作成者によって作成されたと推測されるスプログであった．他のホストでも同様な傾向が見られた．

グサイトからアフィリエイト ID の抽出を行う．ここで，この方式は，主要な ASP (Affiliate Service Provider) のうち 10 社のアフィリエイト ID を抽出する，というものであり，そのカバレージは，約 68 万ブログサイト中で 2 万 5,000 ブログサイトからアフィリエイト ID が抽出できる (抽出率 4%弱) という程度のものである．ただし，この抽出率は，図 2 における「スプログ率 (ランダム)」に相当する割合でスプログ・非スプログが分布している場合に相当する．

ここで，実際に，前節の分析において，同一スパマーの判定を行ったスプログ集合を対象として，アフィリエイト ID の抽出を行った．そして，以下に定義する「アフィリエイト ID 抽出率」の評価を行った．まず,「アフィリエイト ID 抽出率 (ブログサイト単位)」として，既知スプログ s および s に類似するブログサイト集合 $B(H,s,k)$ 中の要素のうち，アフィリエイト ID が抽出できたブログサイト (厳密には，スプログ，もしくは，人手によるスプログ・非スプログの区別が未判定のブログサイト) の割合を算出し，これを評価した．

アフィリエイト ID 抽出率 (ブログサイト単位)

$$= \frac{\sum_{s \in SP_{seed}(H,k)} \left(\begin{array}{c} s \text{ および } B(H,s,k) \text{ 中のスプログサイトおよび} \\ \text{未判定ブログサイトの内,アフィリエイト ID を含むサイト数} \end{array} \right)}{\sum_{s \in SP_{seed}(H,k)} \left(\begin{array}{c} \text{集合 } \{s\} \cup B(H,s,k) \text{ 中のスプログサイトおよび} \\ \text{未判定ブログサイト数} \end{array} \right)}$$

また，これとは別に，$SP_{seed}(H)$ 中の既知スプログの数の単位で，以下の「アフィリエイト ID 抽出率 (ブログサイト集合単位)」を算出し，これを評価した．

アフィリエイト ID 抽出率 (ブログサイト集合単位)

$$= \frac{\left| \left\{ s \in SP_{seed}(H,k) \;\middle|\; \begin{array}{c} s \text{ または } B(H,s,k) \text{ 中のスプログサイトおよび未判定} \\ \text{ブログサイト中に，アフィリエイト ID を含むサイトが存在する} \end{array} \right\} \right|}{\left| SP_{seed}(H,k) \right|}$$

これらの評価結果を図 4 に示す．ここでの分析対象のうちの相当数がスプログであることを考慮すると，両者の評価尺度のうち，特に,「アフィリエイト ID 抽出率 (ブログサイト単位)」は極端に低くなっている．この理由としては，以下が挙げられる．

(1) 主要な ASP のうち，比較的利用者の多い R 社および Ah 社のアフィリエイト ID の抽出が十分に実装できていない．

(2) ブログサイトに直接アフィリエイト広告を貼るのではなく，アフィリエイト広告を多数掲載した広告ページを別途作成し，ブログサイトには，その広告ページへのリンクのみを掲載するタイプのスプログが存在する．

以上の結果をふまえると，現状では，アフィリエイトの情報だけに依存して，スプログの収集・検出を行うことは十分ではないと言える．したがって，アフィリエイトの情報を補足する有力な情報の一つとして，スプログの HTML 構造の類似度が有効であると言える．

4. アフィリエイト ID を用いて収集したスプログの分析

本節では，スプログにおいて，アフィリエイト ID の同一性と HTML 構造の類似性の間の相関の有無を分析する．具体的には，複数のスプログにおいて共有されているアフィリエイト ID を用いて収集したスプログ集合を対象として，HTML 構造の類似性の分析を行う．

4.1 アフィリエイト ID およびブログデータセット

本節では，文献 5) の手順にしたがって，まず，2009 年 12 月 1~31 日の期間に，ブログホスト 8 社[⋆1] から RSS を定期的に取得することによって収集したブログ記事から，主要な ASP のうち 10 社のアフィリエイト ID を抽出した．この段階でのブログサイト総数は，ブログホスト 8 社全体で 19,460 サイトであり[⋆2]，本節では，このうち，3 節における分析と同様に，ブログホスト 3 社 (S 社，F 社，および，C 社) のブログサイトにおいて，ASP のうち 10 社のアフィリエイト ID を含むサイト (5,236 サイト (S 社)，6,348 サイト (F 社)，および，1,740 サイト (C 社)) を分析の対象とする．ブログホスト H における分析対象のブログサイト集合を $B_{af}(H)$ と定義する．

ここで，ブログホスト 8 社全体での 19,460 サイトにおいて，10 以上のブログサイトに含まれていることが確認されたアフィリエイト ID は 182 あり，そのうち，スプログに含まれており，スパマーによって使用されていると判定された ID は 173 となった (スプログ率は 95.1%)．以下では，この 173 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

⋆1 S 社，F 社，C 社，J 社，L 社，A 社，Yh 社，W 社．ただし，実際に，アフィリエイト ID が抽出できたのは，このうち，J 社，Yh 社を除く 6 社．

⋆2 実際に，2009 年 12 月 1~31 日の期間に収集されたブログサイト数は，688,666 サイトであり，このうち，主要な ASP として 11 社のアフィリエイトを使用しているブログサイトは，55,507 サイトであった．また，実際にアフィリエイト ID が取得できた ASP は，11 社のうちの 10 社であり，ブログサイト数としては，31,769 サイトであった．このうち，ブログホストが使用するアフィリエイト ID 以外を含むブログサイト数は 25,140 サイトとなる．さらに，2010 年 7 月の時点において，実際にブログサイトのトップページが取得できたブログサイト数は，19,460 サイトとなった．

表 2　アフィリエイト ID 数およびアフィリエイト ID を含むブログサイトの総数

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	93	1,569
F 社	91	2,985
C 社	5	47

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．この 173 のアフィリエイト ID について，各ブログホストに出現したアフィリエイト ID の数，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 2 に示す．

4.2　アフィリエイト ID を共有するスプログにおける HTML 構造の類似性の分析

アフィリエイト ID x を含むスプログ s ($\in SP_{af}(H,x)$) 同士の間で HTML 構造の類似度を測定し，アフィリエイト ID の同一性と HTML 構造の類似性の間の相関の有無を分析した．相関の有無を評価するために，$k = 0.05,\ 0.1,\ 0.15,\ 0.2,\ 0.25,\ 0.3$ の各値に対して，各アフィリエイト ID ごとに，スプログ集合 $SP_{af}(H,x)$ 中で類似するスプログの組が存在するか否かを判定し，以下の「類似スプログ出現率 (アフィリエイト ID 単位)」を算出する．

類似スプログ出現率 (アフィリエイト ID 単位)

$$= \frac{\left\lvert\left\{ \text{アフィリエイト ID } x \;\middle|\; \exists s, s' \in SP_{af}(H,x),\ \ k - 0.05 \ \le\ \mathrm{Rdiff}(s,s') \ \le\ k \right\}\right\rvert}{\text{アフィリエイト ID } x \text{ の総数}}$$

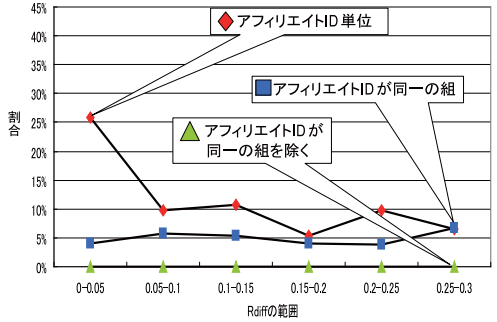
同様に，各アフィリエイト ID ごとに，スプログ集合 $SP_{af}(H,x)$ 中で，類似するスプログがどの程度の割合で含まれるかを測定し，以下の「類似スプログ出現率 (アフィリエイト ID が同一の組)」を算出する．

類似スプログ出現率 (アフィリエイト ID が同一の組)

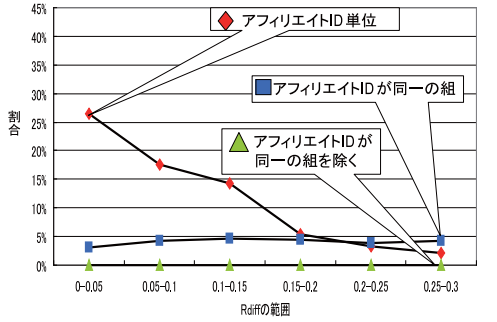
$$= \frac{\displaystyle\sum_x \begin{array}{c}\exists s \in SP_{af}(H,x) \text{ について，} s' \in SP_{af}(H,x) \text{，ただし，} \\ (k - 0.05 \ \le\ \mathrm{Rdiff}(s,s') \ \le\ k) \text{ となる } s' \text{ の最大個数}\end{array}}{\displaystyle\sum_x \left\lvert SP_{af}(H,x)\right\rvert - 1}$$

また，比較対象として，各アフィリエイト ID ごとに，スプログ集合 $SP_{af}(H,x)$ 中のスプログと，分析対象のブログサイト集合 $B_{af}(H)$ のブログサイト (ただし，$SP_{af}(H,x)$ の要素ではない) の間で，類似する組がどの程度の割合で含まれるかを測定し，以下の「類似ブログサイト出現率 (アフィリエイト ID が同一の組を除く)」を算出する．

(a) S 社



(b) F 社



(c) C 社

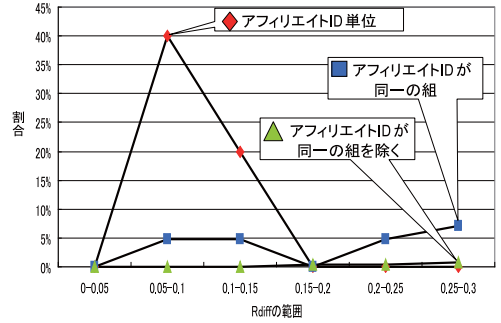


図 5　類似スプログ出現率および類似ブログサイト出現率の評価結果

類似ブログサイト出現率 (アフィリエイト ID が同一の組を除く)

$$= \frac{\displaystyle\sum_x \begin{array}{c}\exists s \in SP_{af}(H,x) \text{ について，} b \in B_{af}(H),\ b \notin SP_{af}(H,x) \text{，ただし，} \\ (k - 0.05 \ \le\ \mathrm{Rdiff}(s,b) \ \le\ k) \text{ となる } b \text{ の最大個数}\end{array}}{\displaystyle\sum_x \left(\left\lvert B_{af}(H)\right\rvert - \left\lvert SP_{af}(H,x)\right\rvert\right)}$$

以上の結果を図 5 に示す．いずれのホストにおいても，アフィリエイト ID が同一のスプログ集合においては，HTML 構造が類似するスプログの組が一定の割合で含まれることが分かる．実際に，分析対象のアフィリエイト ID 数が十分な数含まれる S 社および F 社では，Rdiff の範囲が 0.15 以下で，HTML 構造が類似するスプログの異なり総数が 100 サイト以上となる．また，アフィリエイト ID の個数の単位では，Rdiff の範囲が 0.15 以下となる類似スプログが少なくとも一組存在する ID が 10 個以上存在する．一方，アフィリエイト ID が同一のスプログ以外の組み合わせでは，Rdiff の範囲が 0.15 以下となる組は，数個しかなかった．以上の結果から，アフィリエイト ID が同一のスプログ集合においては，同一のスプログ作成者によって作成された複数のスプログが存在し，それらの HTML 構造が類似している可能性があることがわかる．

一方，同一のアフィリエイト ID を含むにもかかわらず，HTML 構造が類似しないスプログの組も一定数観測された．これらを人手で分析したところ，以下の傾向がみられた．(i) 単なるテキストの貼り付けでなく，多様な HTML タグ構造を持ったメール文面等の貼り付けによって自動生成されたスプログの場合，同一のスパマーが作成しているにもかかわらず，HTML 構造が類似しない．(ii) HTML 文書の形式で作成された複数のパーツを無作為に組み合わせることにより自動生成されたスプログの場合，同一のスパマーが作成しているにもかかわらず，HTML 構造が類似しない．したがって，今後は，アフィリエイト ID が取得できず，しかも，上記のように HTML 構造の類似性の検出が容易でないスプログに対して，HTML 構造の類似性以外の特徴を併用してスプログの特徴を検出する必要がある．

5. おわりに

本論文では，スプログの HTML 構造の類似性およびアフィリエイト ID という異なる二種類の手がかりの特性を分析し，それらの間に相関があることを示した．さらに，既知のスプログに対して HTML 構造が類似するブログサイトを大規模に収集することにより，既知スプログに類似するスプログが高密度で自動収集できた．また，高類似度なスプログに対して，作成者 (スパマー) の同一性の判定を人手で行い，同一のスパマーが作成したと推定されるスプログについても，高密度で収集できたた．本論文の分析結果においては，両者の手がかりは相補的であり，今後は，両者を併用した場合の特性の分析を進める必要がある．特に，現時点において，主要な ASP (Affiliate Service Provider) にわたって，網羅的にアフィリエイト ID を抽出するのは容易ではない．したがって，研究の初期段階においては，相対的にカバレージの大きい HTML 構造の類似度が主導することにより，同一スパマーによって作成されたスプログの候補を収集し，そこから，リンク先の広告サイトまでを含めて，広義のアフィリエイト情報を抽出する技術を確立することが不可欠である．

参 考 文 献

1) 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.
2) Z.Gyöngyi and H.Garcia-Molina. Web spam taxonomy. In *Proc. 1st AIRWeb*, pp. 39–47, 2005.
3) 原正憲, 長谷巧, 山本匠, 山田明, 西垣正勝. スパムブログとアフィリエイトの関連性に関する一考察. 情報処理学会論文誌, Vol.50, No.12, pp. 3206–3210, 2009.
4) 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. In *WebDB2008*. 情報処理学会, 2008.
5) 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. In *WebDB2010*. 情報処理学会, 2010.
6) 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. DEIM フォーラム論文集, 2010.
7) 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.
8) P.Kolari, T.Finin, and A.Joshi. Spam in blogs and social media. In *Tutorial at ICWSM*, 2007.
9) P.Kolari, A.Joshi, and T.Finin. Characterizing the splogosphere. In *Proc. 3rd Ann. Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2006.
10) 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.
11) 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.
12) G.Mishne, D.Carmel, and R.Lempel. Blocking blog spam with language model disagreement. In *Proc. 1st AIRWeb*, 2005.
13) 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.
14) 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.