アフィリエイトIDを用いたスパムブログ収集手法

石井　聡一† 　福原　知宏†† 　増田　英孝† 　中川　裕志†††

† 東京電機大学大学院未来科学研究科　〒101–8457 東京都千代田区神田錦町 2–2
†† 産業技術総合研究所サービス工学研究センター　〒135–0064 東京都江東区青海 2-3-26
††† 東京大学情報基盤センター　〒113–0033 東京都文京区本郷 7–5–1
E-mail: †ishii@cdl.im.dendai.ac.jp, ††tomohiro.fukuhara@aist.go.jp, †††masuda@im.dendai.ac.jp, ††††n3@dl.itc.u-tokyo.ac.jp

あらまし　本論文では，アフィリエイトプログラムの収入を目的としたスパムブログ (スプログ) を分析するために，HTML テキスト中のハイパーリンクなどに含まれるアフィリエイト ID に着目し，アフィリエイト ID を用いたスプログ収集手法を提案する．本研究では，スプログ作成者は同一のアフィリエイト ID を用いてスプログを大量に生成している，との仮説を立て，アフィリエイト ID を単位として分析することにより，複数のブログサービスにまたがってスプログを大量に生成しているスパマの特定が可能となった．提案手法を用いてスプログの分析を行った結果，同一のアフィリエイト ID でブログサイトを 10 以上管理しているスパムアフィリエイト ID は調査期間中に抽出したアフィリエイト ID の 0.8%と少ないが，そのスパムアフィリエイト ID が出現するブログサイト数は調査期間中に抽出したアフィリエイト ID を含むブログサイト全体の 22.4%に上ることがわかった．また，10 ブログサイト以上に出現する 173 のスパムアフィリエイト ID を基にアフィリエイト ID の紐付けを行ない，7ヶ月間スパムアフィリエイト ID を継続して観測した結果，15,409 のスプログを収集した．

キーワード　アフィリエイト，スパムブログ

A Method for Collecting Splogs using Affiliate IDs

Soichi ISHII†, Tomohiro FUKUHARA††, Hidetaka MASUDA†, and Hiroshi NAKAGAWA†††

† Graduate School of Science and Technology for Future Life, Tokyo Denki University　Nichikicho 2–2, Chiyoda-ku, Tokyo, 101–8457 Japan
†† Center for Service Research, National Institute of Advanced Industrial Science and Technology　Oume 2–3–26, Koto-ku, Tokyo, 135–0064 Japan
††† Information Technology Center, The University of Tokyo　Hongo 7–5–1, Bunkyo-ku, Tokyo, 113–0033 Japan
E-mail: †ishii@cdl.im.dendai.ac.jp, ††tomohiro.fukuhara@aist.go.jp, †††masuda@im.dendai.ac.jp, ††††n3@dl.itc.u-tokyo.ac.jp

1. はじめに

近年，ブログサイトなどに企業の広告を掲載する成功報酬型広告 (以下，アフィリエイトプログラム) が盛んに行われている．文献 [1] によると，日本のアフィリエイト総市場規模は 2008 年度に 813 億 1,000 万円となり，2010 年度には 1,000 億円を超えると予測されている．個人が手軽に Web 上で紹介したい商品の情報を掲載できるようになり，消費者は今までのようなメーカや販売店からの情報だけでなく，購入者の商品レビューなどを簡単に得られるようになった．

一方で，広告収入を目的としたスパムブログ (以下，スプログ) が増加している [2] [3] [4]．スプログとは，広告主への誘導や特定サイトの被リンク数増加を目的とし，機械的に大量に生成されるブログサイトである [6]．このことから，スプログは情報検索品質の低下，ネットワーク資源の浪費といった問題を引き起こす要因となっている [6]．

本論文では，アフィリエイトを利用するスプログ分析の為に，アフィリエイト ID を用いたスプログの収集・分析手法を提案する．本研究では，スプログ作成者 (以下，スパマ) は同一のアフィリエイト ID を用いてスプログを大量に生成している，と

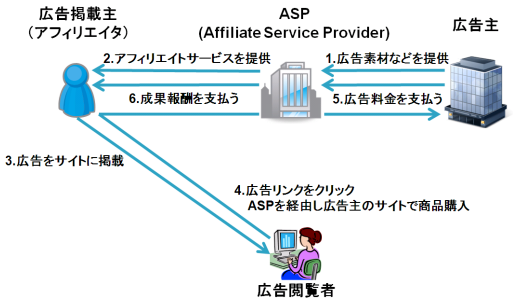

図 1　アフィリエイト利用時のイメージ

図 2　アフィリエイト ID 抽出例

の仮説を立て，アフィリエイト ID を単位として分析することにより，複数のブログサービスにまたがってスプログを大量に生成しているスパマを特定する．実験データとして，アフィリエイトプログラムを利用しているブログサイトを用いる．分析を行った結果，複数のブログサービスにまたがるスパマの特定が可能となった．

本論文の構成は次の通りである．2. でアフィリエイトプログラムの概要，先行研究と我々の提案するアフィリエイト ID に着目した分析アプローチを示す．3. で提案手法であるアフィリエイト ID 特定分析について説明する．4. で提案手法を用いたスプログの収集と分析結果を示す．5. で考察を述べ，6. で本論文のまとめと今後の課題について述べる．

2.　先 行 研 究

本節では，アフィリエイトプログラムの概要について述べ，先行研究との比較を行う．

2.1　アフィリエイトプログラムの概要

アフィリエイトプログラムとは，広告掲載主 (アフィリエイタ) が自身の Web サイトなどに企業サイトへのリンク (アフィリエイトリンク) を貼り，閲覧者がそのリンクを経由し企業サイトで商品の購入などを行うと，アフィリエイタに企業から成功報酬が支払われる広告手法である．図 1 にアフィリエイトプログラムの一般的な利用イメージを示す．

アフィリエイトプログラムの関係者として以下の 4 者が存在する．

（1）　広告主
（2）　ASP(Affiliate Service Provider)
（3）　広告掲載主（アフィリエイタ）
（4）　広告閲覧者

ASP とは，広告主とアフィリエイタの間にある中間業者である．広告主が直接アフィリエイトプログラムを提供することは少なく，広告主に変わり ASP がアフィリエイトプログラムの提供，アフィリエイタへの成功報酬の支払いなどを行っている．

閲覧者がクリックする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 ID(以下，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，どのアフィリエイタを経由してどの広告主のサイトでいくらの商品が売れたのかが，ASP で集計処理されている (3.1 参照)．

2.2　先 行 研 究

スプログ研究では，SVM を用いたスプログ検知 [8] や，リンク解析に着目したスプログフィルタリング [7] などが報告されている．これらの研究では，複数の特徴を用いる必要や，複数のブログサービスにまたがるスパマの分析が難しい．

アフィリエイトプログラムに着目した研究として，原ら [5] の研究では，ブログサイト内に含まれるアフィリエイトリンク数に着目し，1 つのブログサイト内に含まれるアフィリエイトリンク数が多いほどスプログである傾向が高く，アフィリエイトブログの半数がスパムであると報告している．また Wang ら [3] の分析では，スパムサイトから広告主に至るまでのトラフィック分析を行い，スパムが集中するプロバイダの報告をしている．

2.3　本研究におけるアプローチ

我々は，アフィリエイトを利用するスプログの定量分析を目的とし，アフィリエイト ID を用いたスプログ収集・分析手法を提案する．

本研究では，スパマは複数のブログサービスに跨って同一のアフィリエイト ID を含むスプログを大量に作成する，との仮説を立てる．スプログは機械的に大量のブログサイトを生成していると考えられ，また，アフィリエイトプログラムでは，同一アカウントで継続して広告収入を得ているアフィリエイタの報酬単価が増加するからである．

先行研究では，単独のブログサイトや Web サイトを単位として分析し，スパムが集中する ASP，プロバイダなどを報告している [3] [5]．これに対し，提案手法では，アフィリエイトブログに含まれるアフィリエイト ID を用いてアフィリエイタを特定する．これにより，アフィリエイタがどのブログサイトで広告活動を行っているのかという情報を得ることができ，同一のアフィリエイト ID を用いて，複数のブログサービスにまたがってスプログを生成しているスパマの特定が可能となる．

3.　提 案 手 法

提案手法では，ブログサイトからアフィリエイト ID の抽出を行う．また，1 人で複数のアフィリエイト ID を取得するユーザを特定する為，アフィリエイト ID 紐付けアルゴリズムを適用する．

3.1　アフィリエイト ID 抽出方法

図 2 にアフィリエイト ID 抽出例を示す．アフィリエイトリンクには，その広告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 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 の抽出を行う．ASP 各社におけ

groupAffiliateIDs()

Require: *id*

Ensure: *GroupID*

1: $GroupID \leftarrow id$
2: $BlogSites \leftarrow findBlogSites(id)$
3: **for** each $b \in BlogSites$ **do**
4: $IDs \leftarrow findAffiliateIDs(b)$
5: **for** each $cid \in IDs$ **do**
6: **if** $cid \notin GroupID$ **then**
7: $GroupID \leftarrow GroupID \cup groupAffiliateIDs(cid)$
8: **end if**
9: **end for**
10: **end for**
11: **return** *GroupID*

図 3　アフィリエイト ID 紐付けアルゴリズム

表 1　2009 年 12 月から 2010 年 6 月までのブログサイト数と ASP11 社でのアフィリエイトブログ数

期間 (1ヶ月)	全ブログサイト数	アフィリエイトブログ数
2009/12	688,666	55,507(8.1%)
2010/01	708,991	56,070(7.9%)
2010/02	711,160	53,814(7.6%)
2010/03	727,732	52,500(7.2%)
2010/04	750,592	53,131(7.1%)
2010/05	694,171	50,795(7.3%)
2010/06	825,786	56,548(6.9%)
合計	2,245,562	181,755(8.1%)

るアフィリエイトリンク中でのアフィリエイト ID を示すパラメータ名の特定は，予備調査により決定した．

3.2　アフィリエイト ID の紐付け方法

図 2 にアフィリエイト ID 紐付けアルゴリズム *groupAffiliateIDs()* を示す．複数のアフィリエイト ID を取得するユーザのブログサイトには，1 つのブログサイトに複数のアフィリエイト ID が含まれる場合がある．本アルゴリズムの入力は紐付けの基となるアフィリエイト ID(*id*) であり，出力は *id* と紐付けられた ID 集合 *GroupID* である．アフィリエイト ID の紐付けではまず，*id* を含むブログサイト集合を返す関数 *findBlogSites(id)* を用いて，紐付け基となる *id* を含む全てのブログサイトを抽出する (2 行目)．次に，抽出した各ブログサイト *b* について関数 *findAffiliateIDs(b)* を用いてアフィリエイト ID の抽出を行い (3-4 行目)，各 ID*cid* について新たなアフィリエイト ID を抽出した場合は，そのアフィリエイト ID も同一ユーザのアフィリエイト ID として紐付けを行う．次に，*cid* を基に再帰的にアフィリエイト ID を探索する (7 行目)．こうして *id* と紐付けられた ID 集合 *GroupID* が最終的な出力として得られる．

4.　アフィリエイト ID を用いたスプログ収集

本節では，提案手法を用いたスプログ収集・分析結果について述べる．

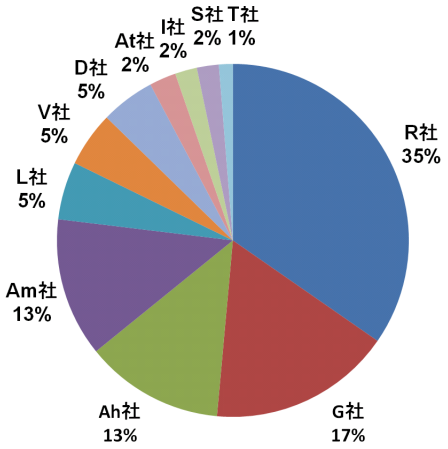

図 4　収集したアフィリエイトブログでの ASP 利用割合

4.1　ブログサイトデータ

分析データとして，ブログサービス運営会社 8 社(注1)から取得したブログサイトを利用する．取得したブログサイトから，利用者の多い ASP を参考に [10] に，ASP11 社(注2)を選定し，ASP11 社のアフィリエイトリンクを含むブログサイトをアフィリエイトブログとし，実験データとして用いる．取得期間は 2009 年 12 月 1 日から 2010 年 6 月 30 日までの 7ヶ月間である．

表 1 に取得したブログサイト数を示す．取得した全ブログが 2,245,562 サイト，その内，ASP11 社のサービスを利用しているアフィリエイトブログが 181,755 サイト (8.1%) となった．

4.2　アフィリエイトブログデータ

取得したアフィリエイトブログから，アフィリエイト ID の抽出を行う．ASP11 社中 2 社のアフィリエイト ID の抽出が困難であり，1 社 (R 社) はある条件時にのみ抽出可能，1 社 (Ah 社) は抽出不可能であった．

図 4 に収集した全アフィリエイトブログ中の ASP 利用割合，表 2 に個人のアフィリエイト ID を抽出したブログサイト数(注3)とアフィリエイト ID 数を示す(注4)．図 4 から R 社が 35%と一番多く，上位 4 社 (R 社，G 社，Ah 社，Am 社) で 78%を占めていることがわかる．表 2 では，全アフィリエイトブログの 48.3%からアフィリエイト ID が抽出できたことがわかる．残り 51.7%のブログサイトでアフィリエイト ID を抽出できなかった原因は，アフィリエイト ID の抽出が困難な ASP2 社 (R 社，Ah 社) の利用者が多く，また，分析対象としていないブログサービス運営会社が利用するアフィリエイト ID も影響している．本研究では，表 2 でのアフィリエイト ID 抽出ブログ 87,771 サイトを主な実験データとして用いる．これは今回取得した全ブログサイト数の 3.9%にあたる．

4.3　アフィリエイト ID を用いたスプログ収集の為の予備実験

本節では，表 2 のアフィリエイト ID 抽出ブログ 87,771 サイ

(注1)：Ab 社, C 社, F 社, J 社, Ld 社, Ss 社, W 社, Y 社,
(注2)：Ah 社, Am 社, At 社, D 社, G 社, I 社, L 社, R 社, S 社, T 社, V 社
(注3)：ブログサービス運営会社が利用するアフィリエイト ID のみを抽出したブログサイトを除去.
(注4)：以下，表中ではアフィリエイト ID を ID と略す.

表 2　2009 年 12 月～2010 年 6 月間でのアフィリエイト ID 抽出ブログサイト数とアフィリエイト ID 数

期間 (1ヶ月)	アフィリエイトブログ数	ID 抽出ブログサイト数	ID 数
2009/12	55,507	25,140	**20,635**
2010/01	56,070	26,064	21,903
2010/02	53,814	25,565	21,392
2010/03	52,500	25,350	21,139
2010/04	53,131	25,259	20,930
2010/05	50,795	25,288	19,784
2010/06	56,548	29,344	23,114
合計	181,755	87,771(48.3%)	68,915

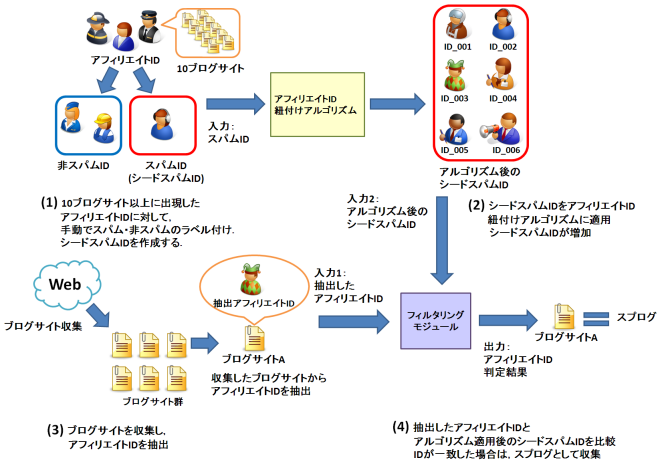

図 5　アフィリエイト ID を用いたスプログ収集手順

トのデータを用いて，スプログの収集と分析の為の 3 つの予備実験を行う．

図 5 にアフィリエイト ID を用いたスプログ収集手順を示す．(1)，(2) について以下の調査を行う．まず，初めに複数のブログサイトに出現する同一のアフィリエイト ID がどの程度あるのか分析を行う．次に，何サイト以上に出現するアフィリエイト ID のスパム率が高いのか調査する．最後にスパムと判定したアフィリエイト ID を用いてアフィリエイト ID の紐付けを行い，紐付けの有効性の調査を行う．

4.3.1　予備実験 1：複数のブログサイトに出現するアフィリエイト ID 数調査

分析期間として表 2 の 2009 年 12 月 1 日～31 日の 1ヶ月間に抽出した 20,635 のアフィリエイト ID に対して，いくつのブログサイトから抽出できたのかを調べる．これにより，複数のブログサイトに出現する同一のアフィリエイト ID がどの程度存在するかがわかる．

表 3 に各アフィリエイト ID のブログサイト出現数を示す．表 3 から 1 つのブログサイトにのみ出現したアフィリエイト ID が 18,946(91.8%) と大多数を占め，2 つ以上のブログサイトに出現したアフィリエイト ID が 1,689(8.2%)，10 サイト以上に出現したアフィリエイト ID は 182(0.9%) と少数であることがわかった．

表 3　2009 年 12 月の各アフィリエイト ID のブログサイト出現数

ブログサイト出現数	ID 数
1 サイト	18,946(91.8%)
2 サイト	961(4.7%)
3 サイト	253(1.2%)
4 サイト	114(0.6%)
5 サイト	52(0.3%)
6 サイト	45(0.2%)
7 サイト	37(0.3%)
8 サイト	24(0.1%)
9 サイト	21(0.1%)
10 サイト	10(0.05%)
2 サイト以上	1,689 (8.2%)
10 サイト以上	**182(0.9%)**
100 サイト以上	14 (0.07%)

4.3.2　予備実験 2：ブログサイト出現数毎でのアフィリエイト ID のスパム率調査

何ブログサイト以上に出現するアフィリエイト ID のスパム率が高いのかを調べるために，出現したサイト数毎でのスパム判定を行う．スパム率は以下の式で算出する．

$$\text{スパム率} = \frac{\text{スパムと判定したアフィリエイト ID 数}}{\text{確認したアフィリエイト ID 数}} \quad (1)$$

スパム判定に関しては，第一著者が目視でアフィリエイト ID ごとのブログサイトを確認し，アフィリエイト ID にスパム，非スパムのラベル付けを行う．本研究ではスパムの判定基準として，文献 [9] のスプログ分類を参考に，アフィリエイトを含み，かつ以下の 4 つの条件のいずれかに当てはまるものをスプログとした．

（1） 広告のみの記事 (オリジナルコンテンツがない)
（2） コピー&ペースト記事
（3） マルチポスト記事
（4） アダルト記事

5 サイト以上に出現したアフィリエイト ID は全てのアフィリエイト ID を確認し，4 サイト以下に出現したアフィリエイト ID はランダムサンプリングで，各 20ID を確認した．

図 6 にブログサイト出現数毎でのアフィリエイト ID スパム率を示す．図 6 から，1 サイトに出現するアフィリエイト ID が大多数 (91.8%) を占めることがわかり，複数のブログサイトに出現するアフィリエイト ID は少数 (8.2%) であった．また，4 サイト以上に出現したアフィリエイト ID では，スパム率が 90%以上であることがわかる．4 サイトに出現したアフィリエイト ID のサンプリング数が 20ID と少ないので，5 サイト以上に出現したアフィリエイト ID についてスパム判定を行うと，スパム率は 94.2%，10 サイト以上に出現したアフィリエイト ID では，スパム率は 95.1%になることがわかった．

表 4 にスパムと判定したアフィリエイト ID 群が含まれていたブログサイト数を示す．表 4 から，同一のアフィリエイト ID を用いて 10 サイト以上のブログサイトに広告を掲載するアフィリエイト ID は全体の 0.9%(182/20,635) と少ない

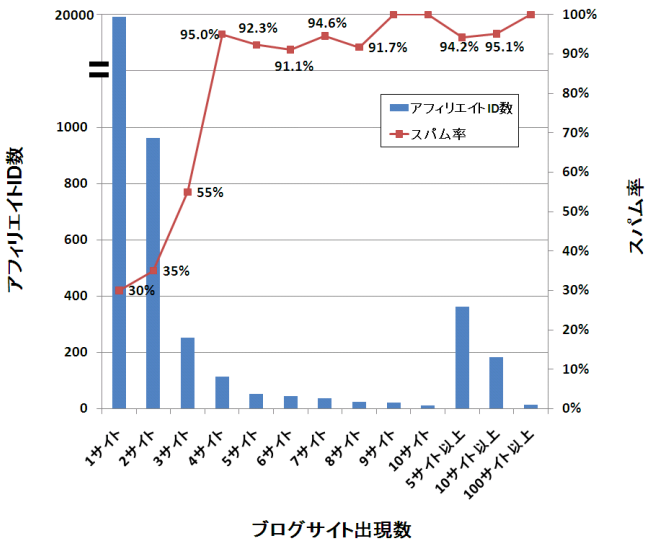

図 6　ブログサイト出現数毎でのアフィリエイト ID スパム率

表 4　スパムアフィリエイト ID 群でのブログサイト数

ブログサイト出現数	ID 数	スパム ID 数	スプログ数
5 サイト以上	361	340	7,589
10 サイト以上	182	**173**	**5,621**
100 サイト以上	14	14	3,119

表 5　173 のスパム ID を基にしたアフィリエイト ID の紐付け

ID の紐付け	ブログサイト数	ID 数	スパム率
紐付け前	5,621	173	95.1%
紐付け後	6,040	475	96.0%

が，その中の 95.1%(173/182) がスパムアフィリエイト ID であり，全体の 0.8%(173/20,635) であった．173 のスパムアフィリエイト ID が含まれていたブログサイト数は 5,621 サイトであり，2009 年 12 月のアフィリエイト ID 抽出ブログサイトの 22.4%(5,621/25,140) に上ることがわかった．

4.3.3　予備実験 3:アフィリエイト ID の紐付けの有効性の調査

表 4 の 10 サイト以上に出現した 173 のスパムアフィリエイト ID を用いて，アフィリエイト ID の紐付けを行う．

表 5 に 173 のスパムアフィリエイト ID を基にしたアフィリエイト ID の紐付け結果を示す．紐付けを行った結果，アフィリエイト ID は 173 から 475 に，ブログサイトは 5,621 から 6,040 に増加した．新たに 302 のアフィリエイト ID を抽出し，419 のブログサイトを収集した．新たに収集したブログサイトをランダムサンプリングで 50 サイト確認したところ 96.0%がスプログであった．このことから，アフィリエイト ID の紐付けがスプログの収集数を増やすことに対して有効なことがわかった．

4.4　アフィリエイト ID を用いたスプログ収集

本節では，4.3.3 で得られた 475 のスパムアフィリエイト ID をシードスパム ID として，2010 年 1 月〜6 月の 6ヶ月間でアフィリエイト ID の紐付けを行わない場合と，行う場合とでのスプログ収集数の比較実験を行う．

表 6　アフィリエイト ID 紐付け前でのブログサイト収集結果

期間 (1ヶ月)	シードスパム ID 残数	ブログサイト数	新出ブログサイト数	スパム率
2010/01	276(58.1%)	5,594	2,385	93.3%
2010/02	258(54.3%)	5,131	1,675	100%
2010/03	230(48.4%)	4,093	973	100%
2010/04	204(42.9%)	3,934	1,011	100%
2010/05	235(49.5%)	5,018	1,095	100%
2010/06	204(42.9%)	5,893	1,272	100%
合計	318	12,906	8,411	98.9%

表 7　アフィリエイト ID 紐付け後でのブログサイト収集結果

期間 (1ヶ月)	シードスパム ID 数	ブログサイト数	新出ブログサイト数	スパム率
2010/01	475	5,912	2,650	100%
2010/02	589	5,331	1,788	96.7%
2010/03	707	4,388	1,069	93.3%
2010/04	777	4,320	1,200	96.7%
2010/05	846	5,436	1,259	93.3%
2010/06	946	6,452	1,403	83.3%
合計	1,206	13,864	9,369	93.9%

4.4.1　アフィリエイト ID の紐付けを行わないスプログ収集実験

本節では，6ヶ月間に 475ID の内どれだけのシードスパム ID が抽出できるのかを調べる．また，各月毎に抽出できたシードスパム ID を基にブログサイトの収集を行い，新たなブログサイトがどれだけ出現するのかを確認する．新たに出現したブログサイトについてはランダムサンプリングで各月 30 サイトのスパム判定を行った．

表 6 にアフィリエイト ID 紐付け前でのブログサイト収集結果を示す．表 6 から，シードスパム ID 数は概ね月を重ねるごとに減少傾向にあった．シードスパム ID を用いることによって，6ヶ月間で，新たに 8,411 のブログサイトが出現し，スパム率は 98.9%であった．

4.4.2　アフィリエイト ID の紐付けを行ったスプログ収集実験

本節では，まず始めにシードスパム ID を基にアフィリエイト ID の紐付けを行い，紐付けられたアフィリエイト ID を用いてブログサイトの収集を行う．アフィリエイト ID の紐付けによって，新たに抽出されたアフィリエイト ID は各月でシードスパム ID に加え，新たに収集されたブログサイトについてはランダムサンプリングで各月 30 サイトのスパム判定を行った．

表 7 にアフィリエイト ID 紐付け後でのブログサイト収集結果を示す．表 6，表 7 から，アフィリエイト ID の紐付けを行った方が新出ブログサイト数が増加することがわかった (月平均 159.7 増加)．また，表 7 からアフィリエイト ID の紐付けにより，シードスパム ID 数が毎月増加していることがわかる (月平均 121.8 増加)．スパム率はアフィリエイト ID の紐付け前よりも劣るが，93.9%となった．アフィリエイト ID の紐付けを行わない場合，シードスパム ID は減少傾向にあり，アフィリエイト ID の紐付けを行った場合，毎月新たなアフィリエイト ID

表 8 処 理 時 間

記事数	容量 (MByte)	処理時間 (sec)	1 記事当たりの処理時間 (sec)	スプログ数
10,000	344	151	0.015	821 (8.2%)
25,000	850	518	0.021	1,882 (7.5%)
50,000	1,701	1,093	0.022	4,038 (8.1%)
100,000	3,447	2,193	0.022	7,942 (7.9%)

を抽出していることからスパマは定期的にアフィリエイト ID を変更していると考えられる．

6ヶ月間に収集したユニークなブログサイト数が 13,864 であることから，67.6%のサイトが新たに確認されたブログサイトとなる(注1)．今回収集したアフィリエイト ID 抽出ブログサイト数が 87,771，アフィリエイト ID 数が 68,915 である．2009 年 12 月の 10 ブログサイト以上に出現する 173(収集したアフィリエイト ID の 0.25%) のアフィリエイト ID にスパムのラベル付けを行うことにより，2009 年 12 月〜2010 年 6 月の 7ヶ月間に 15,409(アフィリエイト ID 抽出ブログサイトの 17.6%) のスプログを収集することができた．

4.5 処 理 速 度

本節では，表 7 のシードスパム ID 1,206 ID を用いて提案手法の処理速度を調べる．本実験では，シードとなるスパム ID を用意した状態で，アフィリエイトブログ記事のアフィリエイト ID 抽出からスプログ検知までの処理時間を調査する．

表 8 に処理時間結果を示す．実験では，記事数毎での全処理時間，1 ファイル当たりの処理時間，収集したスプログ数を計測した．表 8 から，10 万件分のアフィリエイトブログ記事からアフィリエイト ID 抽出とシードスパム ID を比較し，スプログを収集するまでに 2,193 秒 (37 分) で実行できたことがわかる．また，1 記事当たりの処理時間に着目すると，記事数が増加していったとしても処理時間がほとんど変化していないことがわかる．

類似した先行研究が確認できなかった為，参考として竹田ら [9] の研究では，スプログのコピーコンテンツに着目し，25Mbyte 程度のコピー検知コーパスを用いて，21,668 件分のブログ記事のスプログ判定の処理時間を計測している．この研究では，suffix array の構築からスプログ検知までの一連の処理に 3 時間程度を要している．

以上から提案手法は，アフィリエイト ID を含むスプログに限定されるが，高速かつ高精度で動作するスプログ検知手法だと言える．

5. 考 察

本研究では，スパマは複数のブログサービスに跨って，同一のアフィリエイト ID を含むスプログを大量に作成する，との仮説を立てた．その結果，4 サイト以上に出現する同一のアフィリエイト ID のスパム率は 90%以上であり，10 サイト以上ではスパム率は 95.1%に上ることがわかった．スパムアフィリエイト ID を 7ヶ月間継続的に観測することによって，15,409 のスプログを収集することを示した．これらの結果から，本研究での仮説が正しかったことがわかる．

本研究で収集したスプログは，複数のブログサービスに跨ったスプログを同一著者のサイトとして分析が可能である．また，アフィリエイトブログにはアフィリエイト ID だけでなく，商品 ID が含まれている場合があり，先行研究では，スプログによく用いられる単語が確認されている [8]．このことから，収集したスプログから商品情報を抽出し，スパマが好む商品が存在するのか分析を行う必要がある．

6. おわりに

本論文では，アフィリエイトを含んだスプログ分析を目的とし，アフィリエイト ID を用いたスプログの収集・分析手法を提案した．始めにシードとなるスパムアフィリエイト ID を手動で作成する必要があるが，その後はアフィリエイト ID の紐付けを行うことにより，シードスパム ID が増加していくことを観察できた．この結果から，アフィリエイト ID に着目することにより，半自動的にスプログを収集することが可能となった．また，提案手法の処理速度の調査により，分析データが増加していった場合においても，処理速度に大きな変化がないことがわかった．

アフィリエイト ID に着目することにより，ブログサービスをまたいだ横断的な分析が可能となり，高い適合率でスプログを収集することが可能となった．しかしながら本論文で対象としたデータは取得した全ブログサイトの 3.9%とごく少数である．今後，分析対象 ASP を増やすなどして，適用範囲を広げることが必要である．

(注1)：現在の分析では，新たに出現したブログサイトが新しく作られたブログサイトであるのか，記事の投稿をしていなかったブログサイトであるのかの判断を行っていない．

文 献

[1] 矢野経済研究所. アフィリエイト市場に関する調査結果 2009. 矢野経済研究所, 2009.

[2] Z. Gyongyi, and H. Garcia-Molina. Web Spam Taxonomy. In Proc. 1st AIRWeb, pp.39-47, 2005.

[3] Y.-M. Wang, M. Ma, Y. Niu, and H. Chen. Spam double-funnel: Connecting web spammers with advertisers. In Proc. 16th WWW, pp.291-300, 2007.

[4] P. Kolari, T. Finin, and A. Java. Characterizing the Splogosphere, In Proc. 3rd Annual Workshop on Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2006.

[5] 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.

[6] 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.

[7] 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.

[8] P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi. Detecting spam blogs: A machine learning approach. In Proc. 21st AAAI, pp.1351-1356, 2006.

[9] 竹田隆治, 高須淳宏. 複数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), Vol.2, No.1, pp.93-103, 2009.

[10] アフィリエイトマーケティング協会. アフィリエイトプログラムに関する意識調査 2007 年版. アフィリエイトマーケティング協会, 2007.