社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2009－DPS－138（17）
2009－CSEC－ 44（17）
2009／3／5

アフィリエイトに着目したスパムブログ評価方式に関する検討

長谷 巧[1] 山本 匠[2,3] 原 正憲[4] 山田 明[4] 西垣 正勝[2,5]

[1]静岡大学大学院情報学研究科 〒432-8011 浜松市中区城北 3-5-1
[2]静岡大学創造科学技術大学院 〒432-8011 浜松市中区城北 3-5-1
[3]日本学術振興会特別研究，DC1
[4]株式会社 KDDI 研究所 〒356-8502 埼玉県ふじみ野市大原 2-1-15
[5]独立行政法人科学技術振興機構，CREST

あらまし 近年，増え続けるスパムブログが問題となっている．スパムブログを作成する主な目的の１つとしてアフィリエイト収入を得ることが挙げられる．そこで本稿ではブログに含まれるアフィリエイトに着目する．ブログからのアフィリエイトサイトへのリンク数，利用しているアフィリエイトプログラムの種別からスパムブログを判定する方式について検討する．収集したブログに対する調査を通じて本方式の有効性を確認する．
キーワード スパムブログ，アフィリエイト

A study on spam blog detection based on affiliate activity

Takumi NAGAYA[1] Takumi YAMAMOTO[2,3] Masanori HARA[4]
Akira YAMADA[4] Masakatsu NISHIGAKI[2,5]

[1]Graduate School of Informatics, Shizuoka University
[2]Graduate School of Science and Technology, Shizuoka University
[3] Research Fellow of the Japan Society for the Promotion of Science ,DC1
[4]KDDI R&D Laboratories, Inc.
[5]Japan Science Technology and Agency, CREST

Abstract Recently, the number of spam blogs has been dramatically increasing, causing a serious problem. The one of the main reasons of spam blog increase is attractive affiliate income. Thus most of spam blogs have lots of affiliate links to get income. Therefore we propose a technique to detect spam blog by checking the number and/or the sort of affiliate links included in the target blog. This paper carries out a fundamental survey to evaluate the trend of the average number of affiliate links and frequently-used affiliate program in spam blogs.
Keyword spam blog, affiliate

1. はじめに

ブログの普及により，インターネットを通じて多くの人々が簡単に情報発信を行えるようになった．最近ではタレントや政治家，スポーツ選手，その他著名人などによるブログも増加し，ますます注目されている．

しかしその一方でスパムブログが問題となっている．主なスパムブログとしてはニュースサイトや他ブログからの引用やコピー＆ペースト，ワードサラダなどを利用し自動的に作成されたブログ，アフィリエイトの収入を目的としたアフィリエイトリンクアンカーのみを大量に貼り付けたブログ，アダルトサイトや出会い系を目的としたブログなどが挙げられる．スパムブログの増加は，ブログ事業者にとっ

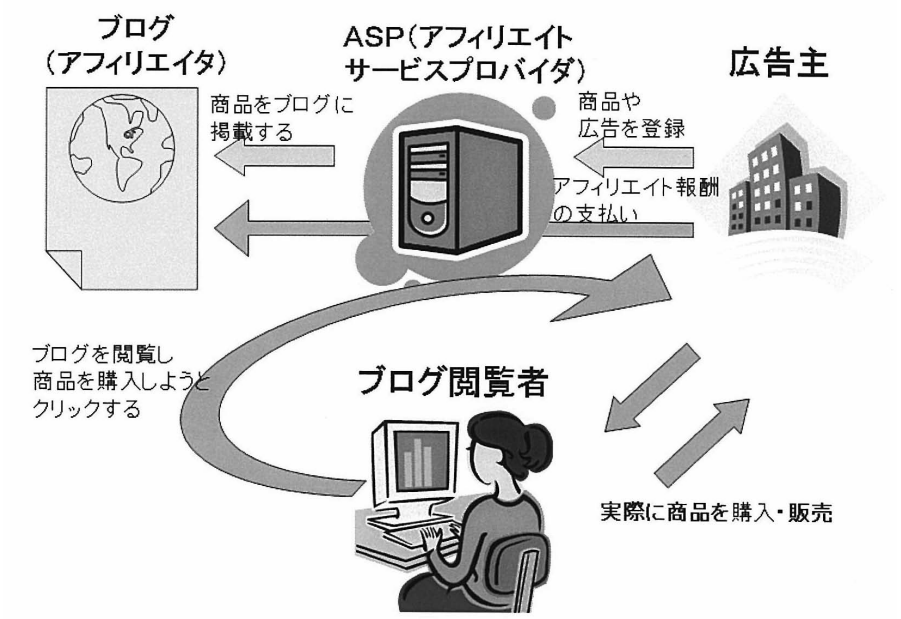

図 1　アフィリエイトの仕組み

ては，大量のブログリソースが消費される，スパムブログ排除のための管理維持費などによるコスト増を引き起こす．またインターネットユーザにとっても，検索エンジンの結果などにスパムブログが掲載されることで，必要な情報へのアクセスがより困難になるなどの問題がある．

2008 年 1 月の総務省の調査において，アクティブブログ（1 カ月に 1 度以上更新するブログ）の 12%がスパムブログと判定☆されている[1]．スパムブログの内容としては，特定のサイトへの誘導を目的としたもの，アフィリエイト収入を目的としたものなどが一般的であった．このため，これらの特徴を検査してやることによってスパムブログを自動判別することができるのではないかと期待できる．

この内，特定サイトへの誘導という特徴に着目したスパムブログを判別する方式としては，文献[2]で検討されている．文献[2]では，ブログ内のリンクアンカーを解析することによって，スパムブログの判別を達成しようとしている．そこで本稿では，アフィリエイトに関する特徴を利用してのスパムブログの判定を試みる．具体的には，ブログに含まれるアフィリエイトリンクアンカーの数および種類からスパムブログを自動判別する方式を提案する．インターネットに実際に存在するブログに対する調査を行い，本方式の有効性について検討する．

2.　アフィリエイトの仕組み

アフィリエイトには，商品を販売している「広告主」，アフィリエイトサービス（以下，アフィリエイトプログラム）を提供している「アフィリエイトサービスプロバイダ（以下，ASP）」，自分のブログの中で商品の紹介をしている「アフィリエイタ」，アフィリエイタのブログを訪問し，そこで紹介されている商品が気に入った場合にはそれを購入する「閲覧者」の４者の登場人物が関わっている．

例えば，成果報酬型のアフィリエイトの流れは図１のようになる[3]．以下詳細な手順につ

☆ 文献[1]の調査におけるスパムブログの判別基準は以下のとおりである．
・機械的に更新している又は他のブログの記事を貼り付けることで更新していると見られるもの．
・ すべてが機械的に更新されているとは見られないが，出来事や関心事等の記述がなく，アフィリエイトや広告記事を大量に掲載しているもの．
・アダルト，出会い系の記事を掲載しているもの．

いて述べる．

1. アフィリエイトによる商品の広告を望む広告主は，ASP のアフィリエイトプログラムに対して商品登録を行う．
2. アフィリエイタは，ASP のアフィリエイトプログラムに対して利用登録を行う．
3. ASP は，アフィリエイトプログラムに利用登録をしたアフィリエイタに，商品登録されている商品の一覧を公開する．アフィリエイタは，自分のブログの中で紹介したい商品を選択する．
4. ASP は，選択された商品に対する商品購入ページの URL（以下，アフィリエイトリンクアンカー）をアフィリエイタに提供する．ここで，アフィリエイトリンクアンカーにはアフィリエイタ，広告主，商品を識別する各 ID などが埋め込まれている．(アフィリエイトリンクアンカーの例：http://○○ASP.com/アフィリエイタ ID+広告主 ID+商品 ID.html)
5. アフィリエイタは，自分のブログの中で商品の紹介し，そのアフィリエイトリンクアンカーを掲載する．
6. アフィリエイタのブログを見た閲覧者がその商品を購入したいと考えた場合，ブログに貼られたアフィリエイトリンクアンカーをクリックする．アンカーに ID が埋め込まれているため，ASP は閲覧者がどのアフィリエイタのブログから商品購入ページに訪れたか知ることができる．
7. 閲覧者が商品を実際に購入した場合，広告主は ASP に報酬を支払い，ASP はアフィリエイタに報酬を支払う．

3. 提案方式

アフィリエイト自体は合法であり，インターネットユーザにとっても自分の欲している商品の情報を知ることができ，また，広告主にとっても効果的な口コミ型の宣伝媒体となっている．ブログにおけるアフィリエイトが問題となるのは，

- 商品情報が掲載されていないブログ：
  アフィリエイトリンクアンカーのみが大量に貼り付けられているブログ
- 不正な商品情報が掲載されているブログ：
  ニュースサイトや他ブログからの引用やコピー&ペースト，ワードサラダなどを利用し自動的に作成された記事に，適当なアフィリエイトリンクアンカーが貼り付けられているブログ
- 公序良俗に反するブログ：
  アダルトサイトや出会い系を目的としたアフィリエイトリンクアンカーが貼り付けられているブログ

などである．そこでまず，正規のブログ（以下，ハムブログ）とスパムブログの間のアフィリエイトに関する傾向の違いを捉える必要がある．

ハムブログにおけるアフィリエイトでは，多くの場合，ブログ内の記事の中でアフィリエイタ自身が 1 つの商品毎に商品を推薦する紹介文を書いてアフィリエイトリンクアンカーと共に掲載している．そのため，1 つの記事内に 1 つのアフィリエイトリンクアンカーが貼り付けられているケースが大勢を占める．一方スパムブログでは，記事内に複数のアフィリエイトリンクアンカーが貼り付けられている場合が多い．よって，ブログ記事内に含まれるアフィリエイトリンクアンカーの数はハムブログとスパムブログを切り分ける 1 つの指標になる可能性がある．

また，上述のとおり，ハムブログにおけるアフィリエイトでは，アフィリエイタ自身が商品の紹介文を書いているため，記事の内容が質・量ともに豊富であることが多い．一方スパムブログでは，商品の紹介文がないものや，適当な他サイトの記事などをコピー&ペーストしただけの紹介文を掲載しているケースが多い．これはスパムブログが自動的・半自動的に作成されていることに起因する．よって，ブログに含まれる記事の内容はハムブログとスパムブログを切り分ける 1 つの指標になる可能性がある．

さらに，ASP ごとにアフィリエイトプログラムで扱う商品ジャンルの傾向が異なるという性質も利用できる．スパムブログにおけるアフィリエイトでは，情報商材（簡単に儲ける方法を教えます等）に関係する商品やアダルト・出会い系の商品を宣伝するものが多いが，これに対し，もっぱらそれらの商品のみを扱っている ASP が存在している．よって，アフィリエ

イタが利用している ASP（アフィリエイトリンクアンカーのリンク先がどの ASP のページにつながっているか）を検査することはハムブログとスパムブログを切り分ける 1 つの指標になる可能性がある．

以上より，ハムブログとスパムブログを切り分ける指標の候補をまとめると次のようになる．

1. ブログの 1 記事内に含まれるアフィリエイトリンクアンカーの数
2. ブログの記事の内容
3. ブログの中で利用されているアフィリエイトプログラムを提供している ASP の種別

この内，指標 2 に対してはブログ記事の構文解析，意味解析が要求されるため，今回，本稿では指標 1 と指標 3 の 2 点に着目する．また，指標 1 については、本来であれば，複数の記事が一つの html ファイルとして表示されている場合には、その中からそれぞれの記事を自動的に抽出する処理が要求されることになるが，今回は各記事の抽出に関する処理については簡略化し、各ブログの html ファイル単位でアンカーの数を計測することとする。

このため、今回のスパムブログ検知は，様々なユーザのブログのトップページ（例：http://blog.co.jp/userID/index.html）の html ファイルを収集の対象とし、html ファイル単位でその中に存在するアフィリエイトリンクアンカーの数をカウントして，一定の閾値以上存在する場合スパムブログと判定する、という手順となる．また，アンカーの数は閾値以下であっても，スパムブログにしか利用されていないと判断される ASP のアフィリエイトリンクアンカーが存在する場合はスパムブログとして判定する．

次章では，本方式の有効性を検討するために，実在のブログを収集し，html ファイルごとのアフィリエイトリンクアンカーの数および ASP の種別とスパムブログとの関係に関して調査する．

4. 調査

4.1. 調査方法

現在，すでに多くの ASP が存在しており，世界中のすべての ASP を把握することは困難である．このため，今回の調査では，収集したブログに含まれているアフィリエイトリンクアンカーを参照し，著者らが確認できた 22 社の ASP（A〜V）を調査対象の ASP とした．

調査対象のブログは，2008 年 9 月 2 日〜2008 年 9 月 19 日にブログ運営会社 X 社のブログの中から無作為に収集した以下の計 1,500 件のブログ（html ファイル）である．

- アフィリエイトブログ：1,000 件
- 非アフィリエイトブログ：500 件

ここで，上記の 22 社の ASP のアフィリエイトリンクアンカーを 1 つ以上含むブログを「アフィリエイトブログ」，22 社の ASP のアンカーを含まないブログを「非アフィリエイトブログ」としている．

収集された 1,500 件のブログの各々に対し，ブログ内に含まれる（A〜V の）ASP ごとのアフィリエイトリンクアンカーの数をカウントするとともに，当該ブログがハムブログかスパムブログかの判定をおこなった．ここで，スパム判定においては，文献[1,4,5]を参考にして以下の 3 つをスパムブログと定義し，調査実施者（著者ら）が判断を下した．

- 自動生成型スパムブログ：
  ニュースサイトや他ブログからのコピー&ペースト，ワードサラダ☆やマルチポスト☆☆などを利用し，自動的・半自動的に作成していると思われるブログ．
- アフィリエイト型スパムブログ：
  商品の写真やアフィリエイトリンクアンカーのみを大量に掲載しており，アフィリエイタによるオリジナルの記事や内容がほとんどないブログ．
- アダルト型スパムブログ：
  記事内に卑猥な文章・写真・動画などのアダルトコンテンツや出会い系記事などが掲載されているブログ．又はそのようなサイトの入り口となっているブログ．

☆ コンピュータによって自動生成された，文法上は正しいが意味は通らない文章．コンピュータでは単語を文法上正しく並べることは可能であるが，意味のある文章を自ら構築することは困難であることに起因する．

☆☆ まったく同じ内容の記事を複数回投稿したり，別のブログに投稿したりすること．

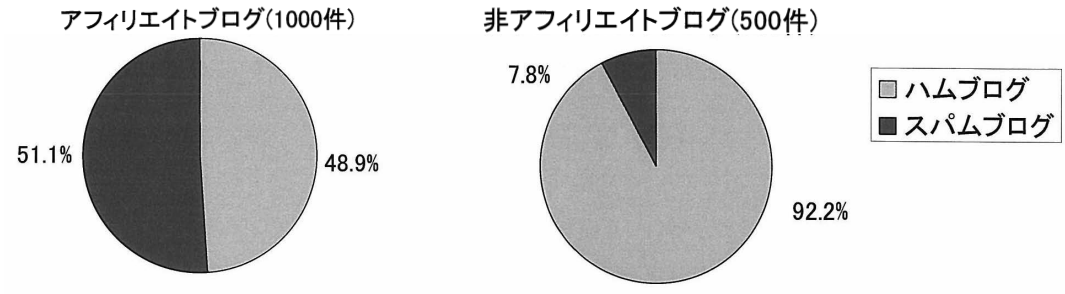

図 2 アフィリエイトとスパムブログの関係

今回は，以下の 2 種類の調査を行った．

- 調査(1)：
  ブログ内に含まれるアフィリエイトリンクアンカーの数（A-V の ASP が提供している各々のアフィリエイトリンクアンカーの数の合計）とスパムブログとの関係性を明らかにする．
- 調査(2)：
  ブログ内で利用されている ASP（A-V）の種別とスパムブログとの関係性を明らかにする．

4.2. 調査結果

調査(1)の結果を図 2 に示す．図 2 はアフィリエイトブログと非アフィリエイトブログにおけるスパムブログの割合を示している．図 2 よりアフィリエイトを行っているブログのほうが，アフィリエイトを行っていないブログよりもスパムブログの割合が高いことが判る．よってアフィリエイトに着目することでスパムブログを判定することは有効であると考えられる．

しかし，アフィリエイトブログにおいてもスパムブログの割合は 5 割程度である．そこで，アフィリエイトブログ 1,000 件をブログ内のアフィリエイトリンクアンカーの合計数によって分類し，分類ごとに，1,000 件中何件のブログがその分類に含まれるか，および，その内の何%がスパムブログであるかを調べた．結果を表 1 に示す．

表 1 より，1 つのブログ内に含まれるアフィリエイトリンクアンカーの数が多いほど，スパムブログである傾向が高いことが判る．特に，100 個以上のアンカーを含むブログについて

表 1 アフィリエイトリンクアンカーの数とスパムブログとの関係

アンカー数	ブログの数	スパム割合
101～	113	88.50%
51～100	95	67.37%
11～50	268	63.81%
6～10	163	55.21%
1～5	361	23.82%
	合計 1,000	平均 51.1%

はスパムブログである可能性が高く，一方，アンカー数が 5 個以下のブログはハムブログである可能性が高い．よって，これらの値は，ハムブログとスパムブログを区別するための閾値になりえる．

次に調査(2)の結果を表 2 に示す．調査(2)では，アフィリエイトブログ 1,000 件をブログ内で利用されている ASP（A～V の 22 社）によって分類し，ASP ごとに，1,000 件中何件のブログがその ASP のアフィリエイトプログラムを利用しているか（1,000 件中何件のブログがその ASP のアフィリエイトリンクアンカーを含むか），および，その内の何%がスパムブログであるかを調べた．

表 2 より ASP ごとにスパムブログの割合が大きく異なることが判る．特に J，L，M，N，O，S 社のアフィリエイトリンクアンカーを含むブログはスパムブログである割合が高い．この結果より，主にスパムブログによって利用されているアフィリエイトプログラムが存在することが確認された．実際に M 社や N 社のアフィリエイトプログラムに登録されている商品を調べたところ，ほとんどの登録商品がアダ

表 2　ASP ごとのスパムブログの割合

ASP	ブログの数	スパム割合
A	282	31.91%
B	444	45.27%
C	167	41.92%
D	50	38.00%
E	185	51.35%
F	54	48.15%
G	51	49.02%
H	25	24.00%
I	1	0.00%
J	144	94.44%
K	23	47.83%
L	74	94.59%
M	22	95.45%
N	2	100.00%
O	6	33.33%
P	32	40.63%
Q	5	80.00%
R	20	35.00%
S	15	93.33%
T	3	66.67%
U	15	26.67%
V	2	50.00%
	合計 1,000	平均 51.1%

ルトグッズであった．よって，これらの ASP のアフィリエイトリンクアンカーの存在の有無によってハムブログとスパムブログを区別する方法は有効であると考えられる．

5.　おわりに

本稿では，アフィリエイトに着目したスパムブログ判別法について検討した．インターネットに実在するブログに対する調査を通じて，本方式の有効性を確認した．

今後は，今回の調査結果を踏まえ，スパムブログ判定の更なる精度向上を達成するために新たな指標の導入を検討する．なお、今回は処理の簡略化のため、3 章で示した指標 2 についてはスパムブログ判定に使用しておらず、また指標 1 についても、複数の記事が一つの html ファイルとして表示されているブログについてはそれを一まとまりとしてアフィリエイトリンクアンカーのカウントを行っている。今後は、指標 2 を導入してのスパムブログ判定、記事当たりアンカー数によるスパムブログ判定についてもその効果を検証していく予定である。また，既存のスパムブログ判定システムに本方式を融合する形で本方式によるスパムブログ検知システムを実装し，大規模な実験を通じて本方式の検知精度を確認していきたいと考えている．

謝辞

株式会社 KDDI 研究所　三宅優様，竹森敬祐様には方式に関しての有益なる助言を頂いた．ここに深く謝意を表する．また，本研究は一部，（財）セコム科学技術振興財団の研究助成を受けている．

参考文献

[1] “ブログの実体に関する調査研究の結果”，総務省，
http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2008/2008-1-02-2.pdf
[2] 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”，データベースと Web 情報システムに関するシンポジウム（DBWeb2008），2008
[3] アフィリエイトで始める!儲かる!ネット通販，宝島社，2004
[4] “ニフティ，スパムブログのフィルタリング技術を開発”，ニフティ株式会社，
http://www.nifty.co.jp/cs/07shimo/detail/080326003337/1.htm
[5] 芳中隆幸，福原知宏，増田英孝，中川裕志，“ブログ空間におけるスパムサイト解析ツールの開発-ユーザ適応型 Splog フィルタリングに向けて-”，暗号と情報セキュリティシンポジウム（SCIS2009），2009