情報処理学会論文誌 Vol. 50 No. 12 3206–3210 (Dec. 2009)

テクニカルノート

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

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

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あげられる．そこで本稿ではブログに含まれるアフィリエイトリンクに着目し，スパムブログを検知する手法を検討するための調査を行った．アフィリエイトリンクを持つ 1,000 件のブログに対して，ブログから各アフィリエイトプログラムへのリンクの数と種類の実態を調べ，アフィリエイトリンクのみに注目を行ったスパムブログ検知について考察した．その結果，含まれているアフィリエイトリンクの数とリンク先のアフィリエイトサービスプロバイダの種類によってスパムブログの特徴を見い出すことができた．

A Study of Relationship between Spam Blogs and Affiliate Marketing

Masanori Hara,[†1] Takumi Nagaya,[†2]
Takumi Yamamoto,[†3,†4] Akira Yamada[†1]
and Masakatsu Nisigaki[†3,†5]

Recently, spam blogs that have been dramatically increasing cause a serious problem. Some spammers aim at getting money by using affiliate. In this paper, we survey the number of affiliate links and frequently-used affiliate program against 1,000 blogs that have affiliate links. We also consider a technique to detect spam blog by checking the number and/or the sort of affiliate links included in the target blogs. As a result, we find the features of spam blogs in the number of affiliate links and the kinds of affiliate service provider.

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．スパムブログとは，広告収入や特定サイトへ誘導することを目的に作成されるブログのことである．スパムブログの増加によって，ブログ事業者には，大量のブログリソースが消費されるだけでなく，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という問題が引き起こされる．またインターネットユーザには，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困難になるなどの問題が引き起こされる．2008 年 1 月の総務省の調査において，308 万件のアクティブブログ（1 カ月に 1 度以上更新するブログ）に対してその 12%の約 37 万件がスパムブログと判定されており[4]，こうした多量のスパムブログを自動検出する技術の確立は社会的にも非常に有用である．

スパムブログの検知には，特徴セットの比較を行い検知する手法[1),2)] やブログの自己相関性を利用した手法[3)] などが提案されているが，これらの手法では事前学習が必要であり，簡便かつ効果的にスパムブログを自動判別する方法が望まれる．上述の総務省の調査で見つかったスパムブログの内容は，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す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．これらの特徴を利用した判定方式のうち，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は文献 5) で検討されている．文献 5) では，ブログ内のリンク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リンクに注目し，アフィリエイトとスパムブログの関係性

†1 KDDI 研究所
KDDI R&D Laboratories Inc.
†2 静岡大学大学院情報学研究科
Graduate School of Informatics, Shizuoka University
†3 静岡大学創造科学技術大学院
Graduate School of Science and Technology, Shizuoka University
†4 日本学術振興会特別研究，DC1
Research Fellow of the Japan Society for the Promotion of Science, DC1
†5 独立行政法人科学技術振興機構，CREST
Japan Science Technology and Agency, CREST

を調査した．調査内容は，ブログトップページのアフィリエイトの有無によるスパムブログ率の違い，アフィリエイトリンクの数とスパムブログ率の関係，各アフィリエイトサービスプロバイダのスパムブログ率の関係である．これらの調査により，ブログトップページに含まれているアフィリエイトリンクの数とリンク先のアフィリエイトサービスプロバイダの種類によってスパムブログの検出に利用可能と考えられる特徴を見い出すことができた．

2. アフィリエイトとスパムブログ

2.1 アフィリエイトの仕組み

アフィリエイトとは，商品を販売する広告主がアフィリエイトサービスプロバイダ（以下，ASP）を通じて出す広告モデルで，アフィリエイタが作成した商品紹介ページを通じて商品を広告し購入を促すものである．ページ内に作られた商品販売ページへのリンクをアフィリエイトリンクといい，アフィリエイトリンクを持つページをアフィリエイトページという．

たとえば，成果報酬型のアフィリエイトの流れは**図 1** のようになっており，アフィリエイタが作成したアフィリエイトリンクを通じて商品が購入された際に，リンクを作成したアフィリエイタへ報酬が支払われるという仕組みである[6)]．

2.2 アフィリエイトを利用したスパムブログ

スパムブログにはさまざまな種類が存在しているが，文献 4)，7)–9) を参考に大きく分けると，アクセス数を稼ぐことを目的としたワードサラダ・コピー&ペースト型スパムブログと金銭取得を目的としたアフィリエイト型スパムブログに分けることができる．なかには 2 つの特性を両方持つスパムブログも存在している．本稿では，特にアフィリエイトとスパムブログの関係に注目をし，調査を行う．

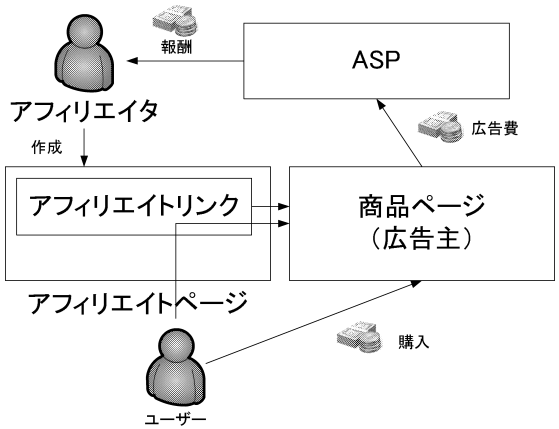

図 1　アフィリエイトの仕組み
Fig. 1　Mechanism of affiliate marketing.

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は自分の欲している商品の評判を知ることができ，広告主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，商品に対する感想がない場合，リンク先の情報をそのまま引用しているだけの場合，公序良俗に反する商品を扱う場合などである．よって，アフィリエイトによるスパムブログ判別を行うためには，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とらえる必要がある．

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 つの商品ごと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とともに掲載しているが，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が貼り付けられている場合が多い[⋆1]．また，なかには情報商材（簡単に儲ける方法を教えますなど）に関係する商品やアダルト・出会い系の商品を主に扱っている ASP が存在している．

以上より，今回の調査では次の点に注目をする．

(1)　ブログトップページに含まれるアフィリエイトリンクの数
(2)　アフィリエイトプログラムを提供している ASP の種別

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，ブログトップページの html ファイルごとのアフィリエイトリンクの数および ASP の種別とスパムブログとの関係に関して調査する．

3. アフィリエイトとスパムブログの関係性調査

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，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アフィリエイトリンクを持つブログ（以下，アフィリエイトブログ）のトップページ 1,000 件とアフィリエイトリンクを持たないブログ（以下，非アフィリエイトブログ）のトップページ 500 件を集めた．これらのブログには 22 社

⋆1 本稿では，検査の簡素化を考慮し，複数の記事から構成されるページ単位（特に，トップページ）でのアフィリエイトリンクの数を計数する．

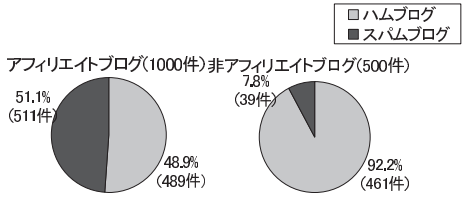

図 2　アフィリエイトとスパムブログの関係
Fig. 2　Relation between affiriate and spam blog.

の ASP（A〜V）へのアフィリエイトリンクが確認された．

3.1　アフィリエイト有無によるスパムブログ率の違い

まずは調査 1 として，アフィリエイトブログ 1,000 件と非アフィリエイトブログ 500 件のスパムブログ率について，当該ブログがハムブログかスパムブログかの判定を目視で行いラベル付けを行った．ここでのスパムブログとは 2.2 節で紹介されたワードサラダ・コピー&ペースト型ブログと金銭取得を目的としたアフィリエイト型ブログのことである．アフィリエイトブログだけに絞らなかったのは，アフィリエイトという特徴だけから，種々のスパムを自動判別できるのかということを調査するためである．

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方が，アフィリエイトを行っていないブログよりも 6 倍程度スパムブログの割合が高いことが分かる．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に含まれるアフィリエイトリンクについての調査を行った．

3.2　アフィリエイトリンクの傾向とスパムブログ率

収集されたアフィリエイトブログのトップページ 1,000 件に対し，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の数をカウントするとともに，ラベル付けされた結果と比較し，以下の 2 種類の調査を行った．

- 調査 2-1：ブログトップページ内に含まれるアフィリエイトリンクの数とスパムブログとの関係．
- 調査 2-2：利用されている ASP（A〜V）の種別とスパムブログとの関係．

調査 2-1 としてアフィリエイトブログのトップページ 1,000 件に対し，各ページ内に含まれるアフィリエイトリンクの数を調べ，そのリンク数ごとに何%がスパムブログであるかを調べた．結果を**図 3** に示す．破線は実測値を表しており，リンク数が少ないものからブログの数を数えていき，その合計が 30 件を超えたリンク数までを 1 つの範囲として，リンク数の平均とスパムブログ率を算出したもので，実線はそのグラフを対数近似したものである．対数近似を選んだ理由は数個程度のリンクを持つブログは多数存在し，リンク数の増加にともない該当数のブログが急減する対数的な要素を持つと考えたためである．対数近似は，$y = c \ln x + b$ で表され c と b を導出した．

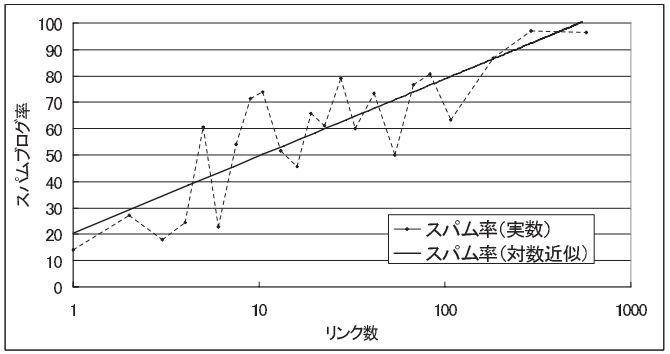

図 3　トップページに含まれるアフィリエイトリンクの数とスパムブログとの関係
Fig. 3　Relation of the number of affriate link and spam blog.

図 3 より，1 つのブログトップページ内に含まれるアフィリエイトリンクの数が多いほど，スパムブログである傾向が高いことが分かる．特に，100 個以上のアンカを含むブログについてはスパムブログである可能性が高く，一方，リンク数が 5 個以下のブログはハムブログである可能性が高い．

次に調査 2-2 の結果を**表 1** に示す．調査 2-2 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）ごとに分類し，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を含むか），および，そのうちの何%がスパムブログであるかを調べた．

表 1 において，合計が 1,000 件を超えているのは複数のアフィリエイトプログラムを利用しているブログが存在しているためである．また，ASP ごとにスパムブログの割合が大きく異なることが分かる．特に E，F，J，O，R，V 社のアフィリエイトリンク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．

表 1　ASP ごとのスパムブログの割合
Table 1　The ratio of spam blog in each ASP.

ASP	ブログの数	スパム割合	ASP	ブログの数	スパム割合
A	472	48.31%	L	26	26.92%
B	288	33.33%	M	23	47.83%
C	191	53.40%	N	21	52.38%
D	170	42.94%	O	15	100.00%
E	144	94.44%	P	15	33.33%
F	94	95.74%	Q	6	50.00%
G	57	50.88%	R	5	80.00%
H	53	52.83%	S	4	75.00%
I	52	40.38%	T	3	66.67%
J	37	97.30%	U	2	50.00%
K	33	45.45%	V	2	100.00%
				合計 1,713	平均 60.8%

3.3 考　　察

調査 1，調査 2-1，調査 2-2 より，スパムブログとアフィリエイトにはある程度の関連が見られる．図 2 より，アフィリエイトリンクがなければハムである確率は高く，図 3 より，アフィリエイトリンク数が多ければスパムである確率が高まることが分かる．以上より，アフィリエイトリンクの有無および数をパラメータとしてスパムブログ検知に組み込むことは有効であると考えられる．なお，ブログトップページに含まれるアフィリエイトリンクの ASP 種数（1 つのブログトップページに何種類の ASP へのリンクを含むか）についても図 3 と同様の調査をしてみたが，ブログトップページ内に 1 種類の ASP へのリンクを含むブログよりも複数の ASP へのリンクを含むブログのほうがスパムブログの確率がやや高まるという程度の結果であった．ただし，ASP へのリンクの数と種類を別々でなく結び付けて考えることにより，よりスパムブログの判定確率を向上させることができる可能性があると考えられる．

また，表 1 より，特定の ASP へのリンクを含むブログがスパムブログである可能性が非常に高くなっていることが分かる．スパムブログのうちスパムブログ率が高かった ASP6 社へのリンクを持つものが 253 件存在した．今回，調査対象としたアフィリエイトブログにおけるスパムブログが 511 件（図 2）であったことを考えると，半数近くのスパムブログがこれらの ASP へのリンクを張っていることが分かる．そのため，この特徴はスパムブログ検知に利用可能であると考えられる．

4. 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 1,000 件に対する調査を通じて，ASP へのリンクを持つスパムブログの傾向を調査した．ASP へのリンクを持つブログは ASP へのリンクを持たないブログに対して 6 倍程度スパムブログが含まれていることが分かった．また，ASP へのリンクの数と種類が多いものほどスパムブログである可能性が高くなり，スパムブログ率が高い ASP が存在することが分かった．これらの特徴を利用することにより，より精度の高いスパムブログ判定が可能になると考えられる．

参 考 文 献

1) Kolari, P., Finin, T. and Joshi, A.: SVMs for the Blogosphere: Blog Identification and Splog Detection, *Proc. 21st National Conference on Artificial Intelligence* (*AAAI 2006*), pp.92–99 (2006).
2) Kolari, P., Java, A., Finin, T., Oates, T. and Joshi, A.: Detecting Spam Blogs: A Machine Learning Approach, *Proc. 21st National Conference on Artificial Intelligence* (*AAAI 2006*), pp.1351–1356 (2006).
3) Lin, Y.-R., Sundaram, H., Chi, Y., Tatemura, J. and Tseng, B.L.: Splog Detection Using Self-similarity Analysis on Blog Temporal Dynamics, *Proc. 3rd international workshop on Adversarial information retrieval on the web* (*AIRWeb 2006*), pp.1–8 (2006).
4) 総務省：ブログの実態に関する調査研究の結果，総務省（オンライン）．入手先 http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2008/2008-1-02-2.pdf（参照 2009-09-16）
5) 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），2B-1 (2008).
6) 宝島社：アフィリエイトで始める! 儲かる! ネット通販，宝島社 (2003).
7) ニフティ株式会社：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．入手先 http://www.nifty.co.jp/cs/07shimo/detail/080326003337/1.htm（参照 2009-09-16）
8) 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），1E1-3 (2009).
9) Kolari, P., Java, A. and Finin, T.: Characterizing the Splogosphere, *3rd Annual Workshop on the Weblogging Ecosystem*: *Aggregation, Analysis and Dynamics*

(*WWE 2006*), 公演番号 6 (2006).

(平成 21 年 7 月 28 日受付)
(平成 21 年 9 月 11 日採録)

原　　正憲（正会員）

1981 年生．2003 年名古屋大学工学部電気電子情報学科卒業．2005 年同大学大学院工学研究科電子情報工学専攻修了．同年 KDDI 株式会社入社後，（株）KDDI 研究所出向，ネットワークセキュリティについての研究に従事．現在に至る．

長谷　　巧（正会員）

2007 年静岡大学情報学部情報科学科卒業．2009 年同大学大学院修士課程修了．同年株式会社デンソー入社．在学中，情報セキュリティに関する研究に従事．

山本　　匠（正会員）

2006 年静岡大学情報学部情報科学科卒業．2007 年 9 月同大学大学院修士課程修了．現在，同創造科学技術大学院博士課程，日本学術振興会特別研究員（DC）．情報セキュリティに関する研究に従事．

山田　　明（正会員）

2001 年神戸大学大学院自然科学研究科電気電子工学専攻博士前期課程修了．同年ケイディディアイ（株）入社．現在，（株）KDDI 研究所ネットワークセキュリティグループ研究主査．2009 年東北大学工学部情報科学研究科博士後期課程修了．暗号プロトコル，インターネットセキュリティの研究に従事．電子情報通信学会，ACM 各会員．

西垣　正勝（正会員）

1990 年静岡大学工学部光電機械工学科卒業．1992 年同大学大学院修士課程修了．1995 年同博士課程修了．日本学術振興会特別研究員（PD）を経て，1996 年静岡大学情報学部助手．1999 年同講師，2001 年同助教授．2006 年より同創造科学技術大学院助教授．2007 年より准教授．博士（工学）．情報セキュリティ，ニューラルネットワーク，回路シミュレーション等に関する研究に従事．