日本語ブログ空間におけるスパムブログ発見手法の提案

寒河江明博† 勝野裕文†
東京電機大学大学院 理工学研究科†

1. はじめに

スパムブログはアフィリエイト収益や商品販売サイトへの誘導を目的として生成されるブログのことで、スパムブロガーは Web 上に存在する多様なコンテンツやサービスを巧みに組み合わせることで、機械的にスパムブログを大量生成している。

Kolari[2,3]は英語ブログ空間を対象に、文書分類に用いられている手法を応用し、機械学習(SVM)によるスパムと非スパムブログの分類を試みている。一方、Lin, et al.[5]は、スパムブログに見られるエントリ間の更新間隔やトピックの類似性に着目したスパム発見手法を提案している。また、石田[6]はブログとスパムキーワードをノードとした2部グラフ上でスパムブログが大規模クラスターとして出現することを見出し、それらクラスターを連鎖的に抽出することでスパムブログを取り出す手法を提案している。

本研究では、日本語ブログ空間において機械学習によるスパム・非スパム分類に有効な特徴量を検討することを目的とする。そこで日本語ブログ空間における[2,3]の手法の有効性を検証し、さらにブログの更新頻度や、リンク、画像といったブログ記事上に潜在的に含まれるメタ情報を用いた手法を試みる。

2. ブログ解析

本研究では、SVM によるスパム検出を行うために、RSS 情報を元にしたブログ解析を行い特徴量の抽出を行う。ここではブログ解析の手法を概説する

ブログの RSS に付随する情報は以下の通りである。(4)~(8)は、投稿日時が新しいエントリ(通常数件~数十件)毎に記載されている。

表 1: ブログ RSS

(1)ブログタイトル, (2)ブログ URL, (3)ブログ概要 (4)エントリタイトル, (5)エントリ URL, (6)著者名 (7)エントリ概要, (8)エントリの投稿日時

(5)を用いて各エントリを入手後、エントリの記事本文を抽出する。これはエントリ上にブログホストが自動配置する記事とは無関係な広告やリンクを排除するためである。本文抽出には事前に用意した各ブログホストが定めている本文タグのリストを用いてそのタグ内容を読み込むか、リストに登録されているタグがない場合は、RSS 内のエントリ概要(<description>の内容)と、対応するエントリページの各 の内容との類似度を計算し、最も高い類似度を含むものを本文として抽出する。本研究では記事本文を主とした各エントリの情報とその RSS からブログの特徴量を評価する。

Methods for Detecting Spam Blogs in Japanese Blogosphere
† Graduate School of Science and Engineering, Tokyo Denki University

3. 実験・評価

DataSet

本実験では、日本国内の主要なブログホスティングサービス 12 サイトから 2008 年 11 月 11 日～12 月 11 日の 30 日間に収集したデータを用いる。データは各サービスが配信する新着ブログ RSS を定期的に取得することで得られたものである。スパム比率は、ブログサービス毎に大きな差があるため、今回の実験ではこれらのサービスのうちスパム率が高い 6 サイトのデータを対象にした。評価に用いるデータは、この中からランダムに約 3000 件サンプリングしたデータで、ラベリング後の内訳は、スパム 1436 件、非スパム 1517 件である。

手順

本研究では、SVM として LibSVM[1]を用い、サンプルデータからランダムに選んだ半数を学習データ、残りをテストデータとして実験を行う。

評価は再現率(Recall)、精度(Precision)、F1 値によって行い、このデータ選出～実験・評価までの過程を、特徴語を利用する手法では 3 回、メタ情報を用いる手法では 10 回繰り返し平均値を最終的な評価値とする。尚、本稿では良結果を示した RBF カーネルを用いた場合のみ示す。

特徴語を利用する手法

Kolari[2,3] の手法を参考に、Bag-of-Words、Anchorsを特徴量とした手法を日本語ブログ空間に適用する。 これらの手法はブログに出現する単語を特徴量とする手法で、事前に特徴量として用いる単語リストを用意する必要がある。特徴量の評価は、スパム・非スパムブログの特徴語リストから、そのリスト上の各単語を次元に持つ特徴ベクトルをブログ毎に作成し、ブログに出現している単語であれば 1、そうでなければ 0 として各次元の値を評価する。特徴語リストは、学習データ中のスパム、非スパムブログそれぞれに対して、Term Frequencyをベースにしたスコアリングを行い、各単語に対して互いのクラス間の相対値を比較し、上位*N*位までの単語を

特徴語として用いる。Bag-of-Words(W)ではこの手法をブログ本文に適用し、Anchors(A)ではアンカーテキストのみを対象とし、それぞれ特徴語リストを作成する。実験結果を図 1に示す。F1 の最高値は以下の通りである。

W:(*N*,Recall,Precision,F1) = (300,0.912,0.903,0.906)
A:(*N*,Recall,Precision,F1) = (500,0.737,0.749,0.743)

また、本実験で得られたスパムブログの特徴語を表 2に示す。

表 2:　スパムブログの特徴語

動画, 無料, 情報, 商品, 離婚, こちら, ダイエット, リンク, サイト, 方法, ネット, 販売, ビジネス, 結婚, 価格, 送料, 人気, 勧め, 投資, 取引, ＦＸ, ブログ, あなた, エロ, すべて, 介護, 相談, 検索, 出会い, 女性, 保険, 詳細, ニュース, 在宅, 広告, パチンコ, 証券, 記事, 紹介, 収入, 関連, 美容, 中古, 最新, アダルト, 簡単, ランキング, 東京, ブランド, 為替, 専門, 効果, 発売, 画像, パソコン, 税込, トレード, ポイント

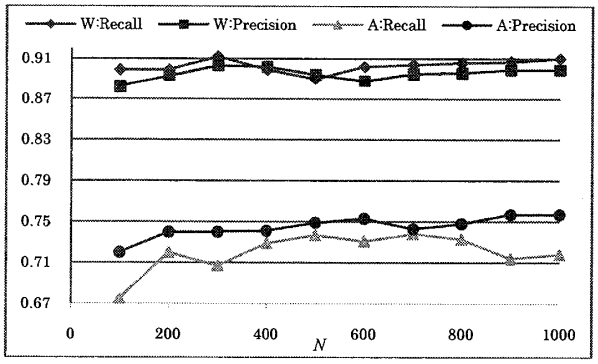

図 1:　Bag-of-Words、Bag-of-Anchors の結果

ブログのメタ情報によるスパム検出

RSSに含まれるメタ情報に加え、本研究ではブログ解析の結果から表 3に示す 15 個の特徴量を抽出し、スパム検出に用いる。

表 3:　ブログのメタ情報として用いる特徴量

1:タイトル長,　**2:**エントリ内文字数,　**3:**ページ容量, **4:**本文長,　**5:**更新頻度,　**6:**本文内外部リンク数, **7:**エントリ内外部リンク数,　**8:**外部リンク URL 平均長, **9:**ユニークリンク率,　**10:**本文内画像リンク数, **11:**エントリ内画像リンク数,　**12:**本文内名詞率, **13:**エントリ内名詞率,**14:**AnchorText 率,**15:**AnchorText 長

この手法では、記事本文以外の情報を一部考慮に入れている。これらの特徴量は、ブログの表面的な特徴であり、ブログに出現する単語等に依存しない。今回は有効と判断したいくつかのパターンに対して行った実験結果を表 4に示す。使用した特徴量は、上記の 1~15 までを 15 ビット列で表現し、用いた場合は 1、用いていない場合は 0 としている。

表 4:　ブログのメタ情報による結果

Feature	Recall	Precision	F1
000000001001100	0.820	0.785	0.802
100010000010110	0.857	0.848	0.852
100111001010111	0.858	0.852	0.855
100011011110110	0.860	0.862	0.861
111111111011111	0.860	0.876	0.868
111111111111111	**0.855**	**0.889**	**0.871**

組み合わせ手法

ここまで挙げた各特徴量を組み合わせて実験を行う。その結果が表 5である。W:Bag-of-Words、A:Bag-of-Anchors、M:メタ、数字は各特徴量の次元数を表す。

表 5:　組み合わせ手法の結果

Feature	Recall	Precision	F1
500A+15M	0.840	0.861	0.850
100W+15M	0.899	0.905	0.902
300W+15M	0.907	0.910	0.909
300W+500A	0.909	0.906	0.907
300W+500A+15M	0.910	0.910	0.910
300W+300A+15M	**0.914**	**0.914**	**0.914**

4.　おわりに

本研究では、日本語ブログ空間を対象に SVM を用いたスパム検出手法において有効な特徴量の検討を行った。結果、各手法で 0.9 前後の F1 値を示すことができた。今後は長期間での有効性を検証しライフサイクルが短いスパムブログに対してどのように本手法を適用していくか検討する必要がある。また有効に作用する特徴量を Lin[5]が用いている Fisher linear discriminant analysis によって適時選択できるようにしたい。

参考文献

[1] C-C. Chang, and C-J. Lin. LIBSVM: A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[2] P. Kolari, T. Finin, and A. Joshi. SVM for the Blogosphere: Blog Identification and Splog Detection. AAAI 2006, 2006.

[3] P. Kolari. Detecting spam blogs: an adaptive online approach. PhD thesis, University of Maryland, 2007.

[4] Y.Sato, et al. Analysing features of Japanese splogs and characteristics of keywords. AIRWeb08, 2008.

[5] Y-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B L. Tseng. Splog Detection Using Self-similarity Analysis on Blog Temporal Dynamics. AIRWeb07, 2007.

[6] 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ダリング. WebDB Forum 2008, 2008.